ALGORITHMES DE CLASSIFICATION

Maurice ROUX
Professeur émérite
Université Paul Cézanne
Marseille, France.

Avertissement

Cet ouvrage a été publié aux éditions Masson, Paris, en 1985. Il est maintenant épuisé et nous mettons en accès libre la présente version électronique, corrigée et améliorée.

La première version de cet ouvrage comportait, à la fin de chaque chapitre des programmes en langage Basic-Applesoft qui sont maintenant obsolètes. Ces programmes ont été convertis en « Visual Basic for Applications » utilisables avec le tableur EXCEL (Microsoft). Ils sont réunis dans le classeur « *AnaDon.xls* » associé à un mode d'emploi inclus dans le fichier « *AnaDon.doc* » lisible avec le traitement de textes WORD (Microsoft). A la fin de chaque chapitre de l'ouvrage figurent les noms des procédures de ce classeur traitées dans le chapitre.

Marseille, Juin 2006.

ALGORITHMES DE CLASSIFICATION

Table des matières

CHAPITRE 1. - Introduction à la classification
1. But de la classification
2. Problèmes et méthodes de la classification automatique
3. Objectifs et plan de l'ouvrage
4. Domaines d'application et points de vocabulaire

CHAPITRE 2. - Exemples de données
1. Psychologie et société (Psysoc)
2. Phytosociologie (Phytos)

CHAPITRE 3. - Préparation des données. Calcul des distances
1. Généralités
    1.1. Données quantitatives ; exemple des causes de décès (Psysoc)
    1.2. Pré-traitement par l'analyse factorielle
    1.3. Variables qualitatives et mixtes
2. Application aux exemples
    2.1. Causes de décès (Psysoc)
    2.2. Phytosociologie (Phytos)
3. Les procédures de calcul de distances

CHAPITRE 4. - La classification ascendante hiérarchique
1. Généralités
    1.1. Principe général des constructions ascendantes
    1.2. Propriétés des formules élémentaires de recalcul
    1.3. Comparaison des agrégations par le saut minimum et par le diamètre
2. Application aux exemples
    2.1. Causes de décès (Psysoc)
    2.2. Phytosociologie (Phytos)
3. Les procédures de constructions ascendantes de hiérarchies

CHAPITRE 5. - Agrégation autour de centres mobiles
1. Principes et problèmes
    1.1. L'algorithme des centres mobiles
    1.2. Moment d'ordre deux d'une partition
    1.3. Avantages et inconvénients de la méthode
2. Application à l'exemple Psysoc
    2.1. Partition en trois classes
    2.2. Partition en quatre classes
3. Les programmes de calcul de centres mobiles

CHAPITRE 6. - Hiérarchie du moment d'ordre deux

1. Principe et problèmes
2. L'algorithme des voisins réciproques
3. Application à l'exemple Psysoc
4. Procédure de calcul

CHAPITRE 7. - Classification descendante hiérarchique

1. Introduction
2. Méthodes basées sur une variable particulière
   2.1. Utilisation de l'une des variables des données
   2.2. Utilisation des variables principales, ou axes factoriels
3. Méthodes basées sur des individus particuliers
   3.1. Sélection d'un point périphérique
   3.2. Sélection de deux points périphériques
   3.3. Sélection de deux points-noyaux
4. Le problème des inversions
5. Application aux exemples
   5.1. Données PSYSOC
   5.2. Données PHYTOS
6. Conclusion
7. Procédure de calcul

CHAPITRE 8. - Aides a l'interprétation

1. Variables quantitatives
   1.1. Interprétation d'une partition
   1.2. Interprétation d'une hiérarchie
2. Variable qualitatives
   2.1. Interprétation d'une partition
   2.2. Interprétation d'une hiérarchie
3. Application aux exemples
   3.1. Données Psysoc (quantitatives)
   3.2. Données Phytos (qualitatives)
4. Les procédures d'aide à l'interprétation

CHAPITRE 9. - Pratique de la classification

1. Choix d'un algorithme
   1.1. Dimensions des données
   1.2. Nature des données
   1.3. Qualité des résultats
   1.4. Temps de calcul
2. Stratégies
   2.1. Hiérarchie puis centres mobiles
   2.2. Centres mobiles suivis d'une hiérarchie
   2.3. Données hétérogènes, emploi de l'analyse factorielle préalable
3. Interprétation des résultats
4. Un programme supplémentaire utile : troncature d'une partition

CHAPITRE 10. - Conclusion

1. Taxinomie de qualité
    1.1. Préparation des données
    1.2. Traitement
    1.3. Interprétation des résultats
2. Classification en tant que pré-traitement
    2.1. Préparation des données
    2.2. Traitement
    2.3. Interprétation

ANNEXE 1. - Les indices de ditances
1. Généralités
2. Cas des données binaires
    2.1. Indices où la présence des attributs joue un rôle prépondérant
    2.2. Indices où les présences et absences d'attributs jouent des rôles équivalents
3. Cas des donnees quantitatives
    3.1. Coefficients de corrélation
    3.2. Mesures de distances
4. Conclusion

ANNEXE 2. - Hiérarchies et ultramétriques
1. Généralités
    1.1. Hiérarchie et ordonnance
    1.2. Hiérarchie indicée et ultramétrique
2. Une ultramétrique particulière la sous-dominante
    2.1. Relation d'ordre sur les métriques
    2.2. Ultramétrique “ sous-dominante ” d'une métrique donnée

BIBLIOGRAPHIE
INDEX

Chapitre 1

Introduction à la classification

1. But de la classification

Comme les autres méthodes de l'Analyse des données, dont elle fait partie, la Classification a pour but d'obtenir une représentation schématique simple d'un tableau rectangulaire de données dont les colonnes, suivant l'usage, sont des descripteurs de l'ensemble des observations, placées en lignes.

L'objectif le plus simple d'une classification est de répartir l'échantillon en groupes d'observations homogènes, chaque groupe étant bien différencié des autres. Le plus souvent, cependant, cet objectif est plus raffiné ; on veut, en général, obtenir des sections à l'intérieur des groupes principaux, puis des subdivisions plus petites de ces sections, et ainsi de suite. En bref, on désire avoir une hiérarchie, c'est à dire une suite de partitions "emboîtées", de plus en plus fines, sur l'ensemble d'observations initial.

Une telle hiérarchie peut avantageusement être résumée par un arbre hiérarchique (figure 1) dont les nœuds (m, n, p, q) symbolisent les diverses subdivisions de l'échantillon ; les éléments de ces subdivisions étant les objets (a, b, c, d, e), placés à l'extrémité inférieure des branches qui leur sont reliées.

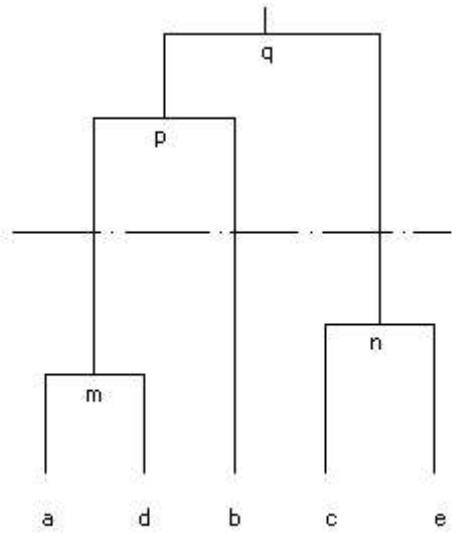


Figure 1*. Exemple d'arbre hiérarchique portant sur cinq objets a, b, c, d, e. Les points m, n, p, q sont les nœuds de l'arbre. Le trait horizontal mixte indique un niveau de troncature définissant une partition en trois classes.*

Le niveau des nœuds, qui est le plus souvent chiffré, est sensé indiquer un degré de ressemblance entre les objets correspondants. Ainsi, sur notre figure 1, les objets a et d se ressemblent plus que les objets c et e. Remarquons, en passant, que si on coupe cet arbre à un niveau intermédiaire entre n et p, on obtient une partition en trois classes de l'ensemble étudié, savoir les parties {a, d}, {b}, {c, e}. En faisant varier ce niveau de troncature on obtient les diverses partitions constituant la hiérarchie.

On voit qu'il ne faut pas confondre classification et classement. Dans un classement on affecte les objets à des groupes préétablis ; c'est le but de l'analyse discriminante que de fixer des règles pour déterminer la classe des objets. La classification est donc, en quelque sorte, le travail préliminaire au classement, savoir la recherche des classes "naturelles" dans le domaine étudié.

2.- Problèmes et méthodes de la classification automatique

Dans cet ouvrage il sera beaucoup question d'algorithmes. Rappelons qu'un algorithme est la description minutieuse de toutes les opérations à effectuer pour obtenir la solution concrète d'un problème. Ainsi on peut parler de l'algorithme permettant de trouver la racine carrée d'un nombre, ou bien pour obtenir le plus grand commun diviseur de deux nombres entiers, etc ...Il ne faut pas confondre algorithme et programme informatique : il peut y avoir plusieurs façons de programmer un même algorithme.

L'un des plus grands classificateurs a, sans aucun doute, été le savant suédois Linné qui, au 18-ème siècle, a établi une classification du monde vivant en général et du règne végétal en particulier, classification encore en vigueur aujourd'hui chez les spécialistes des sciences naturelles. La première moitié du 20-ème siècle a vu un certain nombre de tentatives pour rationaliser le processus mental utilisé par Linné. Mais ce n'est qu'à partir des années 1960, avec la diffusion de l'informatique en milieu universitaire, que sont apparus un grand nombre d'algorithmes automatisant complètement la construction des classifications (Williams and Lambert, 1959, Sokal and Sneath, 1963). Cependant, aujourd'hui encore le support mathématique de ces méthodes reste embryonnaire et ne permet pas d'élire un algorithme aux avantages indiscutables.

Supposons que l'on veuille, par exemple, construire une hiérarchie. L'une des manières de "bien poser" le problème pourrait être de choisir un critère évaluant la fidélité de la représentation hiérarchique au tableau initial des données, et de trouver ensuite un algorithme construisant la hiérarchie la meilleure, au sens de ce critère. Malheureusement on ne sait pas faire cela sauf pour des échantillons très petits, ou pour des critères sans intérêt. La solution qui consiste à examiner l'ensemble de toutes les hiérarchies possibles, pour en retenir la meilleure, se heurte au "mur" de la complexité combinatoire. Le nombre de hiérarchies croît en effet si vite avec le nombre d'objets que, même avec de puissants ordinateurs, il n'est pas réaliste de vouloir les envisager toutes. C'est pourquoi l'on a recours à des heuristiques, c'est à dire des algorithmes dont on considère qu'ils sont suffisamment raisonnables vous donner des résultats satisfaisants.

Grossièrement on peut distinguer trois grands types parmi ces heuristiques. Il y a d'abord les algorithmes construisant une hiérarchie par agrégations successives d'objets, puis de groupes, en fonction des distances entre objets ou groupes. On les appelle "Constructions ascendantes de hiérarchies", en abrégé CAH. A l'inverse les "Constructions descendantes de hiérarchies", en abrégé CDH, procèdent par dichotomies successives. Dans celles-ci l'ensemble tout entier est d'abord scindé en deux, puis chacune de ses parties est, à son tour subdivisée, et ainsi de suite. Dans le troisième groupe de méthodes on peut rassembler toutes celles qui se limitent à l'élaboration d'une partition. Par des algorithmes très divers, ces méthodes ont pour objectif de détecter les zones à forte densité dans l'espace des observations.

Etant donné la faiblesse des bases théoriques de tous ces algorithmes usuels, il serait imprudent de se fier totalement aux résultats ainsi obtenus. C'est pourquoi nous recommandons vivement à l'utilisateur de toujours confronter ses résultats à ceux d'une analyse factorielle (Benzécri et coll. 1973 b, Bertier et Bouroche 1975, De Lagarde 1983, Fénelon 1981, Foucart 1982, Bouroche et Saporta 1980).

3.- Objectifs et plan de l'ouvrage

Dans les pages qui suivent on se propose de donner les bases mathématiques, les algorithmes et les programmes de calcul pour les principales méthodes de classification. Comme notre intention est de fournir aux praticiens les moyens de comprendre et d'utiliser ces méthodes nous avons basé l'exposé sur deux exemples typiques (décrits au chapitre 2) qui sont traités par tous les algorithmes possibles.

Chaque chapitre comporte l'exposé d'un algorithme et son application à l'un ou l'autre des exemples. On explique ensuite la mise en œuvre du programme correspondant et ses principales caractéristiques en vue d'une adaptation éventuelle. Par souci de clarté les développements théoriques importants sont renvoyés en annexe.

Comme la plupart des méthodes commencent par le calcul de distances, on étudiera d'abord les modalités de ce calcul (chapitre 3). On pourra alors décrire les algorithmes usuels de construction ascendante de hiérarchie (chapitre 4), puis un algorithme, devenu classique, de construction d'une partition (chapitre 5). On envisage ensuite des méthodes moins courantes : la construction ascendante selon la variance des distances (chapitre 6) et une construction descendante hiérarchique (chapitre 7). On termine par des calculs complémentaires facilitant l'interprétation des rêsultats (chapitre 8) et par un chapitre (numéro 9) indiquant quelques règles élémentaires à suivre pour le traitement ces données. En conclusion (chapitre 10) nous résumerons les caractéristiques de chacune des techniques décrites en indiquant nos préférences.

4.- Domaines d'application et points de vocabulaire

La classification a un rôle à jouer dans toutes les sciences et techniques qui font appel à la statistique multidimensionnelle. Citons tout d'abord les sciences biologiques : botanique, zoologie, écologie, ... Ces sciences utilisent également le terme de "taxinomie" pour désigner l'art de la classification. De même les sciences de la terre et des eaux : géologie, pédologie, géographie, étude des pollutions, font grand usage de classifications.

La classification est fort utile également dans les sciences de l'homme : psychologie, sociologie, linguistique, archéologie, histoire, etc ... et dans les techniques dérivées comme les enquêtes d'opinion, le marketing, etc ... Ces dernières emploient parfois les mots de "typologie" et "segmentation" pour désigner la classification, ou l'une de ses innombrables variantes. Citons encore la médecine, l'économie, l'agronomie, et nous en oublions certainement !

Dans toutes ces disciplines la classification peut être employée comme une fin en soi ; mais elle l'est souvent, à juste titre, comme une méthode complémentaire à d'autres méthodes statistiques. Elle peut, en effet, aider efficacement à l'interprétation des graphiques d'analyse factorielle, ou bien déterminer des groupes d'objets homogènes, préalablement à une régression linéaire multiple.

Chapitre 2

Exemples de données

Avant d'aborder les méthodes classificatoires nous présentons deux exemples qui nous serviront tout au long de ce livre.

1.- Psychologie et société (PSYSOC)

Notre premier exemple est tiré du livre de E. Todd : "Le fou et le prolétaire" (1979, annexe 2, p 283). Il s'agit de statistiques concernant, pour différents pays occidentaux, les causes de décès, qui selon Mr Todd, sont caractéristiques de l'état de santé mentale de la société (voir tableau 1, six premières colonnes). Notre objectif sera d'établir une classification des pays en fonction de ces taux de mortalité, calculés pour 100.000 habitants.

Afin de juger du bien fondé des classifications nous donnons ici les résultats de l'Analyse factorielle des correspondances de ce tableau (Tableau 1, colonnes F1, F2 et F3). Les variables étant quantitatives on aurait pu appliquer également l'Analyse en composantes principales. Toutefois l'étude des "profils" des pays réalisée par la première nous paraît mieux adaptée au sujet traité, c'est à dire les taux de mortalité comme indicateurs de maladies sociales (voir chapitre 3 pour un complément de justification). Au demeurant, les "poids" des lignes étant relativement comparables, les résultats des deux types d'analyse factorielle sont assez voisins.

	SUICI	HOMIC	AROUT	AINDU	AAUTR	CIRFO	F1	F2	F3
AUSTRIA	241	16	330	43	363	325	-220	-6	108
FRANCE	156	9	225	10	535	328	-210	-3	-110
PORTUGAL	85	19	349	7	281	345	-369	-257	-65
WGERMANY	210	12	230	21	298	269	-245	17	149
BELGIUM	156	10	260	13	367	144	-7	95	-37
FINLAND	251	26	180	29	387	55	258	270	178
SWEDEN	194	11	151	13	384	122	54	214	58
SWITZERL	225	9	195	26	276	128	-15	212	211
ITALY	54	11	219	19	224	319	-484	-287	-90
NIRELAND	40	136	215	18	320	43	727	-691	48
DENMARK	241	6	168	11	230	107	-21	289	334
ICELAND	101	5	179	23	380	9	328	283	-241
SCOTLAND	82	15	155	18	342	59	215	109	-203
SPAIN	40	4	136	17	237	225	-392	-178	-183
NORWAY	104	6	138	22	346	41	234	250	-176
SIRELAND	38	7	182	32	314	37	242	100	-379
NETHERLA	89	7	169	10	218	47	133	142	-68
ENGLANDW	79	10	130	14	203	36	200	141	-65
USA	121	102	220	26	273	158	253	-447	195

Tableau 1*.- Données PSYSOC avec les résultats de l'Analyse factorielle des Correspondances. Les six premières colonnes contiennent les taux de mortalité de différentes causes violentes de décés dans 19 pays occidentaux, en nombre de décès pour 100 000 habitants. Les trois dernières colonnes (F1, F2 et F3) sont les coordonnées factorielles (multipliées par 1000) des pays sur les trois premiers axes de l'Analyse factorielle des Correspondances.*

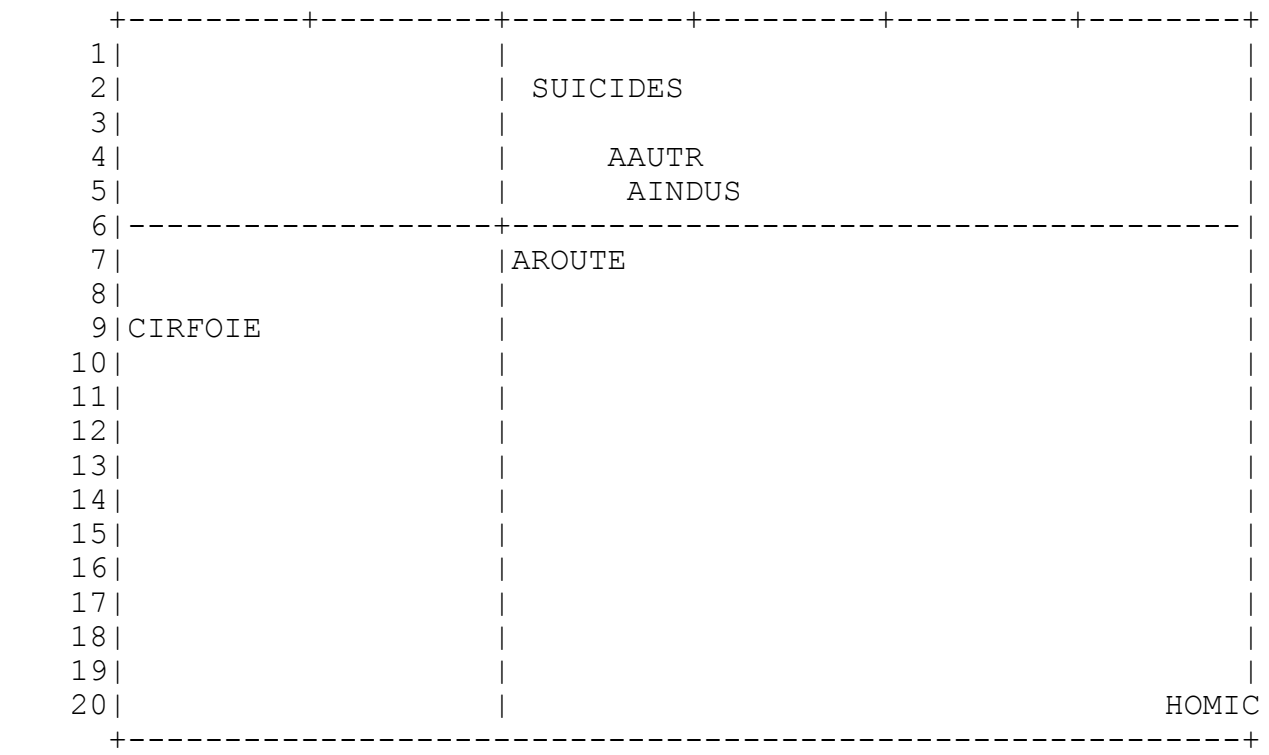


***Figure 1**.- Données PSYSOC, Analyse des correspondances, représentation des variables sur les axes 1 et 2. Ces deux axes expliquent respectivement 44,33 % et 34,41 % de la variance totale.*

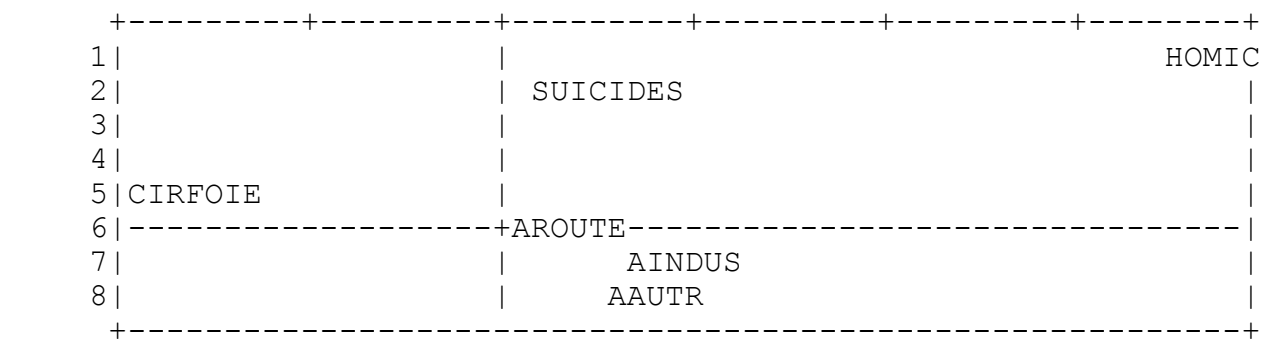


***Figure 1 bis.**- Données PSYSOC, Analyse des correspondances, représentation des variables sur les axes 1 et 3. Ces deux axes expliquent respectivement 44,33 % et 14,96 % de la variance totale.*

Sur le graphique des variables (figure 1) l'axe 1 oppose les homicides aux décès par cirrhose du foie, les différents types d'accidents étant en position intermédiaire. On peut donc interpréter cet axe comme celui de l'agressivité de la société. Le second axe est d'interprétation plus difficile. Outre qu'il temoigne d'un léger effet Guttman (disposition en forme de croissant, *cf* Benzécri 1980, Volle, 1978), il isole principalement les homicides, ceux-ci étant massivement le fait de deux pays seulement l'Irlande du Nord et les USA (figure 2). Enfin le 3-ème axe (figure 1 bis) établit une distinction entre la mort donnée volontairement (suicides et homicides du coté positif de l'axe) et les décès accidentels.

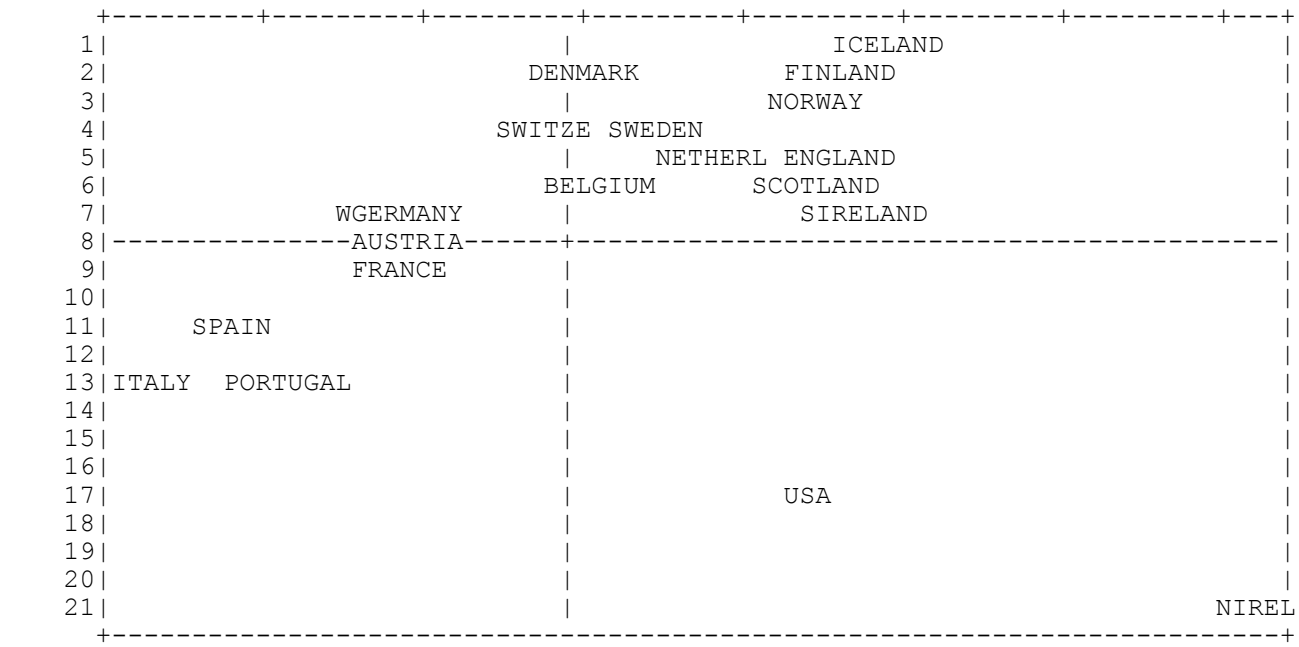


Figure 2.- *Données PSYSOC, Analyse des correspondances, représentation des pays sur les axes 1 et 2. Ces deux axes expliquent respectivement 44,33 % et 34,41 % de la variance totale.*

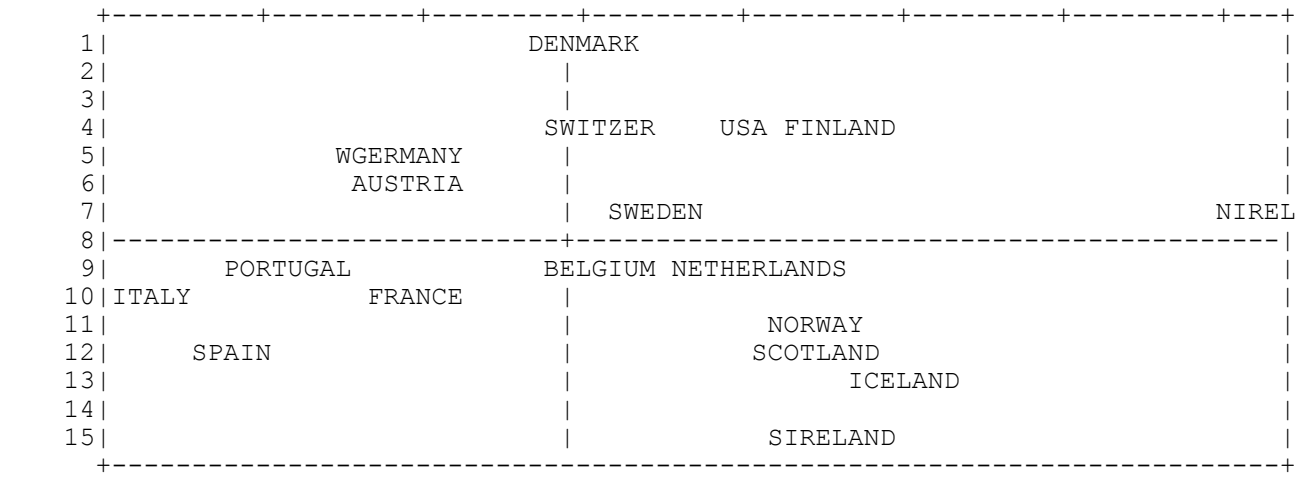


Figure 2 bis.- *Données PSYSOC, Analyse des correspondances, représentation des pays sur les axes 1 et 3. Ces deux axes expliquent respectivement 44,33 % et 14,96 % de la variance totale.*

L'examen du plan 1-2 pour les pays (figure 2) confirme la thèse de Mr Todd sur la similitude entre l'Allemagne et la France du point de vue des tensions internes de la société, alors que l'Angleterre se trouve être plus proche des pays nordiques. On remarque également le regroupement des pays méditerranéens (ESP, PORT, ITAL) dans la zone dominée par la cirrhose du foie ...

2.- Phytosociologie (PHYTOS)

L'étude des affinités de terrain entre espèces végétales porte le nom de phytosociologie. Elle a pour point de départ des enquêtes sur des régions plus ou moins étendues au cours desquelles on effectue des "relevés". Un relevé consiste en la liste des espèces végétales poussant dans un lieu particulier. Le résultat d'une enquête de terrain se met sous la forme d'un tableau rectangulaire où l'usage est de mettre les relevés en colonnes et les espèces en lignes.

		1	1	1	1	1	2	2	2	3	3	3	3	5	5		
	3	4	0	3	4	5	6	3	4	7	0	1	6	8	4	5	
1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	Achillea millefolium
2	0	1	0	0	0	0	1	0	0	1	1	1	1	1	1	1	Agrostis alpina Scop.
5	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	Alchemilla glaberrima Schm.
7	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	0	Alchemilla hybrida L.
10	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	Androsace carnea L.
11	1	1	1	0	1	0	1	0	1	1	0	1	0	1	1	1	Antennaria dioica (L) Gaertn
12	1	1	0	1	1	1	1	1	1	0	1	0	1	0	0	1	Anthoxanthum odoratum L.
20	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Aster alpinus L.
21	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	Astragalus campestris (L) Ten
24	0	0	1	1	1	1	1	1	1	0	1	0	0	0	1	1	Avena versicolor Vill.
26	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	Botrychium lunaria (L) Sw.
29	1	1	0	1	1	1	1	1	0	1	0	0	0	0	0	1	Campanula scheuchzeri Vill.
41	1	1	0	1	1	1	1	0	0	0	0	1	1	0	1	0	Carex sempervirens Vill.
42	1	0	0	1	1	1	0	0	0	0	1	1	1	1	1	1	Cerastium arvense var. strict.
45	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	Cirsium acaule (L) Webb.
48	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	Crepis aurea L.
50	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	Deschampsia flexuosa (L) Trin
53	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	Draba aizoides L.
55	0	0	1	0	0	0	0	0	0	1	0	0	0	1	1	1	Elyna myosuroides (All) Degld
57	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	Erygeron sp.
60	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	Euphrasia minima L.
61	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1	Festuca halleri
62	1	1	1	1	1	1	0	0	0	1	0	1	1	0	1	0	Festuca macrophylla
63	0	1	0	1	0	0	0	1	0	0	0	0	0	0	1	1	Festuca violacea
64	1	0	0	0	1	1	1	0	0	0	0	0	1	0	1	1	Galium pumilum (Lmk) Ry
65	1	1	1	0	1	0	1	0	0	0	1	0	1	0	0	0	Gentiana alpina Vill.
67	1	1	1	1	1	1	0	0	0	1	1	1	1	0	1	1	Gentiana campestris L.
68	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	Gentiana kochiana Per. Song.
69	1	1	1	0	1	0	1	0	0	0	0	0	1	0	1	1	Gentiana nivalis L.
70	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	Gentiana punctata L.
71	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	Gentiana verna L.
72	1	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	Geum montanum L.
75	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	Gregoria vittaliana (L) Duby
77	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	0	Hieracium glaciale (Reyn) Lach.
79	0	1	1	0	1	1	1	1	1	0	0	0	0	0	0	0	Hieracium pilosella L.
82	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	Homogyne alpina (L) Cass.
84	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	Juncus trifidus L.
86	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Leontodon helveticus
87	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	Leontodon pyrenaicus Gouan
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	Lotus corniculatus
95	0	0	0	1	0	1	0	1	0	1	0	0	0	0	1	0	Luzula spicata (L) DC
98	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	Minuarta rupestris (Scop) Sch.
100	1	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	Nardus stricta L.
105	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	Pedicularia rostratospicata
109	0	1	1	0	0	0	1	0	0	1	1	1	0	1	0	0	Phyteuma hemisphericum L.
112	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	Phyteuma orbiculare L.
113	1	1	0	1	1	1	1	1	1	1	0	0	1	0	1	1	Plantago alpina L;
114	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	Poa alpina L.
116	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	Polygonum viviparum L.
117	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	Potentilla aurea L.
120	1	0	0	0	0	0	1	0	0	1	1	1	1	1	0	0	Potentilla grandiflora L.
125	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	Pulsatilla vernalis L.
126	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	Ranunculus pyrenaicus L.
129	0	0	0	1	0	1	0	1	0	0	0	0	0	0	1	0	Sagina glabra (Willd) Fenzl.
130	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Sagina linnaei Presl.
131	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	Salix herbacea L.
144	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0	0	Sempervivum arachnoideum L.
145	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	Sempervivum montanum Jacq.
156	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	Thymus serpillum (L) Lyka
157	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	Trifolium alpinum L.
158	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	Trifolium badium Schreb.
159	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	Trifolium pratense ssp nival
160	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	Trifolium thallii Vill.
163	1	1	1	0	1	0	1	0	1	1	0	1	1	0	0	1	Veronica allionii Vill.
166	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	0	Veronica bellidioides L.
168	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	Veronica serpyllifolia L.

***Tableau 2**.- Données PHYTOS : présence (1) ou absence (0) de 66 espèces végétales dans 16 relevés du Plateau d'Emparis (Hautes-Alpes, France). Les numéros des relevés sont écrits en colonnes, sur les deux premières lignes.*

On porte, à l'intersection de la ligne i et de la colonne j, un 1 si l'espèce i est présente dans le relevé j, et un zéro dans le cas contraire. On note parfois un coefficient d'abondance au lieu de la simple présence-absence ; toutefois, dans notre exemple, nous ne prenons en compte que cette dernière.

Le tableau 2 recense 66 espèces dans un ensemble de 16 relevés. Ces données sont extraites d'un ensemble plus vaste, de 55 relevés, effectués sur le plateau d'Emparis (2200 m d'altitude, Hautes Alpes) par G. Roux, et déjà analysé par ailleurs (*Cf* chapitres Alpes I et II dans Benzécri et coll., 1973 a). Pour réduire la taille du tableau on a, en outre, éliminé une trentaine d'espèces qui n'étaient présentes qu'une seule fois et dont le rôle est donc minime. L'objectif de cette étude est de vérifier le bien fondé de la classification des pelouses "à nard" (du nom de l'espèce dominante) que nous avions obtenue précédemment sans les dissocier des autres relevés. Celle-ci s'établissait ainsi :

Sigles des groupements	Relevés	Noms des groupements
Pan	13, 15, 23	Nardetum alpigenum
Pacn1	3, 4, 14, 16, 24	Festucetum halleri Sunass. Nardetosum
Pacn2	10, 54, 55	Festucetum halleri Subass. Nardetosum Faciès à Elyna et Salix
Pac	27, 30, 31, 36, 38	Festucetum halleri Sensu stricto

Tableau 3.- *Données PHYTOS : partition des 16 relevés en 4 classes appelées groupements.*

Les noms des groupements sont établis en fonction des espèces "caractéristiques". Par exemple, le dernier groupement est appelé Festucetum halleri parce que son espèce caractéristique est *Festuca halleri*. Mais, si chaque espèce, prise individuellement, s'accommode de terrains plus ou moins variés, les associations végétales sont, en général, caractéristiques de conditions d'environnement très précises (*Cf* Guinochet, 1955, 1973)

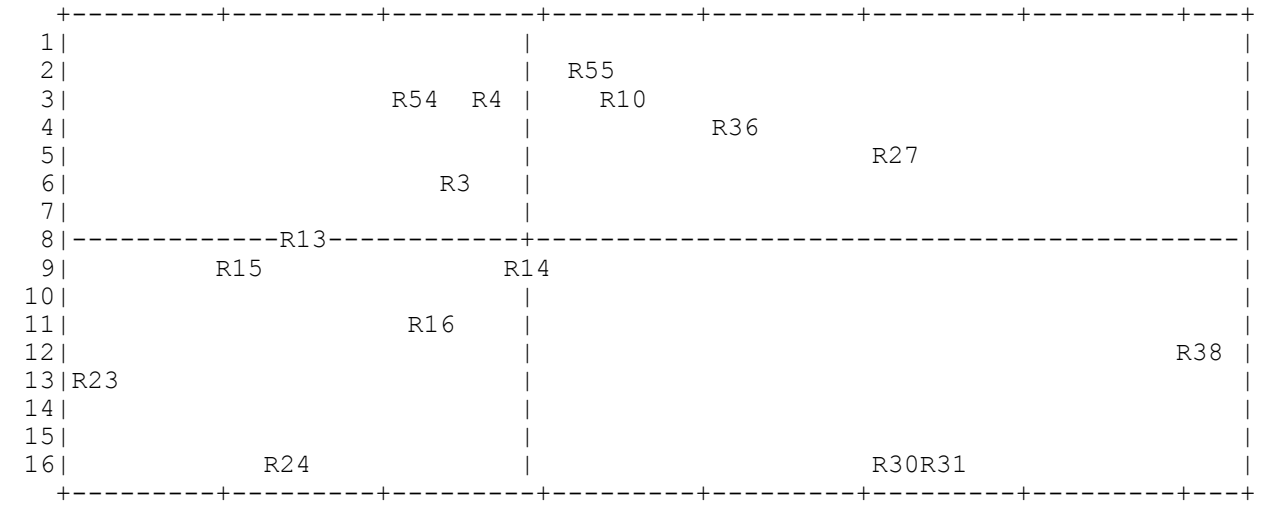


Figure 3.- *Données Phytos, Analyse des correspondances, représentation des relevés sur les axes 1 (horizontal) et 2 (vertical). Ces deux axes expliquent repectivement 21,32 % et14,53% de la variance totale.*

Après Analyse factorielle des correspondances, en examinant conjointement les deux plans factoriels formés des axes 1-2 et 1-3 (figures 3 et 4), on reconnaît l'existence des groupements Pan (13, 15, 23) et Pac (27, 30, 31, 36, 38) aux deux extrémités de l'axe 1. La réalité des deux autres groupements est plus contestable. La classification automatique confirmera-t-elle ou infirmera-t-elle cette partition ?

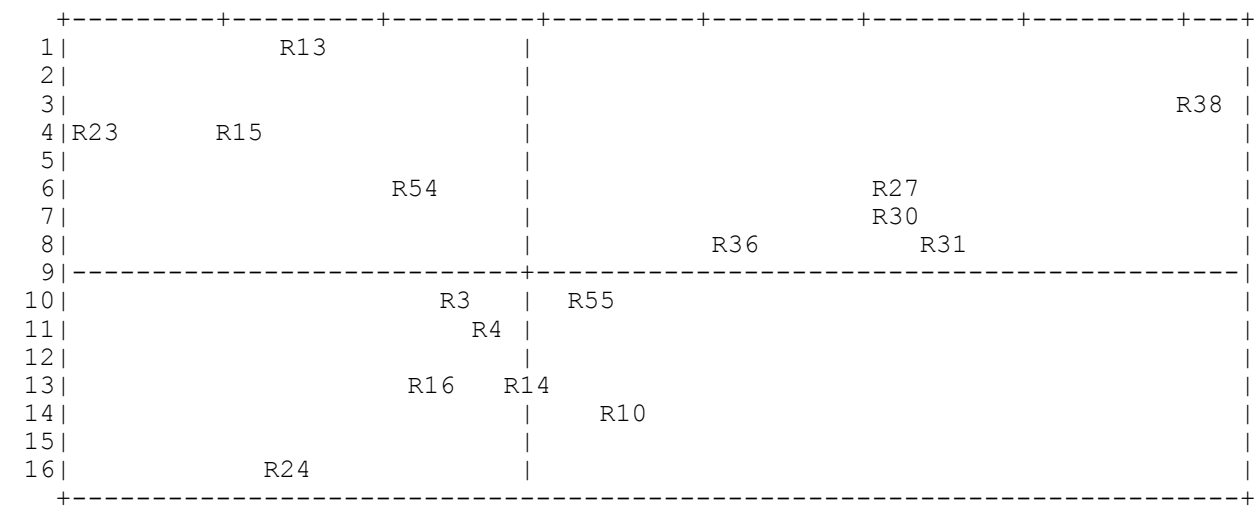


Figure 4.- *Données Phytos, Analyse des correspondances, représentation des relevés sur les axes 1 (horizontal) et 3 (vertical). Ces deux axes expliquent respectivement 21,32 % et 10,64 % de la variance totale.*

Chapitre 3

Préparation des données, calcul des distances

La plupart des algorithmes de classification ont pour point de départ une mesure des distances, ou dissemblances, entre les objets. Or il existe une infinité de façons pour évaluer ces dissemblances, et la formule retenue aura une influence décisive sur les résultats. C'est pourquoi nous croyons que l'utilisateur doit réfléchir consciencieusement sur cette question en fonction de chaque problème pratique. Nous donnons ci-dessous quelques idées générales ; elles sont complétées par des considérations mathématiques plus précises dans l' annexe 1.

1. Généralités

1.1.- Données quantitatives ; exemple des causes de décès (Psysoc)

Dans nos données sur les causes sociales des décès il nous faut commencer par calculer les distances entre les pays. La formule la plus utilisée est celle de la distance euclidienne usuelle :

$$d^2(i, i') = \Sigma_j (x_{ij} - x_{i'j})^2$$

où x_{ij} désigne le nombre de décès de cause j dans le pays i. Par exemple, pour l'Autriche et la France on aura :

$$d^2(AUST, FRAN) = (241-156)^2 + (16-9)^2 + \ldots + (325-328)^2 =$$
$$= 7225 + 49 + 11025 + 1089 + 29584 + 9 =$$
$$= 48981$$
$$d(AUST, FRAN) = 221.3$$

Un premier problème apparaît immédiatement : les nombres qui mesurent les homicides (deuxième terme dans la somme ci-dessus) sont beaucoup plus petits que les autres. Leur contribution à la distance (ici 49) sera donc, en général, beaucoup plus faible que celle des autres colonnes du tableau. Pour rééquilibrer les rôles des variables l'usage est d'opérer leur réduction, c'est à dire de diviser les valeurs par l'écart-type de la variable considérée.

Le second problème provient des différences globales dans les taux de mortalité. Il peut en effet arriver que deux pays aient une répartition des décès analogue, mais que, pour l'un des deux, les quantités soient toujours plus faibles que pour l'autre. Seules sont conservées les proportions entre les catégories de décès. On peut alors considérer que ces deux pays souffrent des mêmes malaises sociaux, l'un à un degré moindre que l'autre. Cependant, comme la distance euclidienne repose sur les écarts absolus, ces deux pays seront vraisemblablement éloignés et donc classés dans des catégories distinctes. On dit qu'il y alors un "effet de taille". On peut pallier cette difficulté en calculant la somme des décès par pays, puis en remplaçant chaque valeur par son rapport à cette somme.

Mais cette transformation ne résout pas tous les problèmes. En effet si plusieurs variables sont liées au même phénomène sous-jacent, elles seront corrélées entre elles et apporteront plusieurs fois la même information. Pour éviter cet inconvénient on peut utiliser une formule de distance particulière appelée "métrique du khi-deux" qui fait intervenir à la fois les poids x_i des lignes et x_j des colonnes. Ces poids ne sont autres que les sommes des termes de la ligne i ou de la colonne j :

$$d^2(i, i') = \Sigma_j \, (1/ x_{.j}) \, \{x_{ij}/ x_{i.} - x_{i'j}/x_{i'.}\}^2 \qquad (1)$$

Les termes de chaque ligne i sont rapportés à leur somme $x_{i.}$. Une variable j contribue à la distance en raison inverse de son poids $x_{.j}$. Une autre solution intéressante s'offre à nous que nous allons examiner en détail ci-dessous.

1.2.- Pré-traitement par l'Analyse factorielle

Cette opération consiste à effectuer avant la classification, soit une Analyse en composantes principales (ACP), soit une Analyse factorielle des correspondances (AFC), selon ce qui parait le mieux adapté aux données et aux objectifs poursuivis. On prend alors, comme nouvelles données pour la classification, les coordonnées des objets sur les premiers axes factoriels obtenus, c'est à dire ceux qui apportent le plus d'information (cf Benzécri 1980, Foucart 1982, Volle 1978, etc ...). Bien qu'il implique beaucoup de calculs, ce détour vaut la peine d'être fait car il présente de nombreux avantages :

1)Le plus important d'entre eux est que l'Analyse factorielle fournit des nouvelles variables non correlées entre elles et élimine donc la dernière difficulté examinée ci-dessus.

2)Le délicat problème du choix de la distance initiale se trouve également résolu : c'est la distance euclidienne usuelle qui s'impose. En effet, si l'on a opté pour l'ACP, elle redonne approximativement la distance euclidienne usuelle que l'on aurait pu calculer sur les données brutes ; si l'on a opté pour l'AFC, la distance euclidienne usuelle sur les facteurs est à peu près égale à la métrique du Khi-deux sur les données brutes. Dans les deux cas le degré d'approximation est d'autant meilleur qu'on travaille sur un plus grand nombre de facteurs. Bien entendu il ne s'agit pas d'une méthode miracle ! Le choix de la distance se trouve remplacé par le choix du codage préalable des données en vue de l'analyse factorielle. Mais les différents codages possibles sont maintenant bien connus et éprouvés. (*Cf* Benzécri 1980, Roux et Guittonneau, 1977).

3) L'Analyse factorielle des correspondances surmonte élégamment le problème de l'effet de taille et permet de traiter des données très hétérogènes, par découpages en classes de valeurs des variables quantitatives, et mise sous forme disjonctive complète de l'ensemble des variables.

4) On y gagne également sur le plan informatique. Comme on ne conserve rarement plus de cinq à dix facteurs le tableau des données est d'une taille raisonnable et peut, en général, tenir dans la mémoire centrale de l'ordinateur. D'ou un gain de temps et une plus grande facilité de programmation. Mais, surtout, on n'a qu'un seul programme de distance à programmer : celui de la distance euclidienne.

5) Les facteurs de l'analyse factorielle sont très stables - c'est à dire que de petites erreurs de mesures, ou bien la suppression d'observations douteuses, ne modifient quasiment pas les coordonnées sur les axes, ni, par conséquent les classifications calculées d'après ces coordonnées. Or c'est précisément un défaut fréquent de ces méthodes que d'être sensibles à de petites fluctuations des données. Dans l'analyse factorielle celles-ci modifient surtout les derniers facteurs, c'est à dire ceux que l'on ne prend pas en compte dans notre stratégie.

6) L'analyse factorielle permet une autre approche des données et facilite l'interprétation des classifications obtenues.

La seule difficulté de cette méthode réside dans le choix du nombre d'axes factoriels à prendre en considération. Toutefois l'utilisateur sera guidé dans ce choix par l'examen des décroissances successives des pourcentages d'inertie des axes factoriels. Il faut arrêter lorsque celles-ci deviennent négligeables. D'autre part un autre critère important est de ne conserver que les facteurs que l'on arrive à interpréter.

1.3.- Variables qualitatives et mixtes

Lorsque les variables sont qualitatives la stratégie ci-dessus s'applique encore, avec cette restriction que seule l'analyse des correspondances est justifiée sur le plan mathématique. Il convient pour cela de mettre les données sous forme disjonctive complète. C'est à dire qu'à chaque état de variable, ou modalité, on fait correspondre une colonne du tableau final. En regard d'une observation, occupant une ligne du tableau, on met un "1" dans les colonnes indiquant ses qualités et des zéros partout ailleurs (*cf* Benzécri 1980, Foucart 1982, ...).

Toutefois pour certaines données où les variables sont à deux modalités - présence ou absence d'un attribut - il arrive que l'absence n'ait pas la mème valeur significative que la présence. Il est alors préférable de coder chaque attribut sur une seule colonne (au lieu de deux) avec un "1" si l'attribut est présent et un zéro s'il est absent. C'est le cas en phytosociologie (cf exemple 2 au chapitre précédent) où la présence d'une plante est une indication plus importante que son absence relativement à la nature du sol, au climat, etc ...

De nombreux chercheurs ont d'ailleurs mis au point des formules de distances prenant en compte cette remarque. Ainsi l'indice de Jaccard fournit généralement un bon point de départ pour une classification. Cet indice est basé sur le nombre c d'attributs communs (c'est le nombre d'espéces présentes simultanément dans deux relevés de plantes) et sur les nombres p et q d'attributs possédés par chacune des deux observations considérées :

$$d = 1 - c/(p + q - c) \qquad (2)$$

Le dénominateur de la fraction représente le nombre d'attributs existant soit dans l'une, soit dans l'autre , soit dans les deux observations. Cet indice vaut zéro lorsque les deux observations sont tout à fait identiques, et un lorsqu'elles n'ont aucun attribut en commun. Primitivement cet indice a été créé comme une mesure de ressemblance :

$$s = c/(p + q - c) \qquad (3)$$

La ressemblance vaut zéro quand les deux observations n'ont pas de caractères communs et un lorsqu'elles sont identiques. Mais nous préférons l'expression sous forme de distance, qui permet de n'avoir qu'un seul programme de classification pour travailler sur des données qualitatives ou quantitatives. De nombreuses formules analogues sont données en Annexe 1 avec les remarques qu'elles nécessitent.

Enfin dans le cas où les données contiennent un mélange de variables qualitatives et quantitatives, il est encore possible de combiner des formules pour obtenir une expression de la distance entre observations (voir annexe 1). Mais cette manière de faire comporte tellement d'arbitraire qu'il vaut mieux, dans ce cas, découper les variables quantitatives en classes de valeurs, que l'on considère ensuite comme des modalités. On applique alors l'AFC puis la classification sur les coordonnées factorielles.

2.- Application aux exemples

2.1.- Causes de décès (PSYSOC)

Les données sur les causes des décès, déjà examinées ci-dessus (paragraphe 1.1) sont constituées de valeurs additives : la somme des nombres d'une ligne du tableau représente, en effet, pour un pays, ce que E. Todd appelle le taux de mortalité sociale, c'est à dire le nombre de décès pour 100.000 habitants dus à des causes sociales. La somme des termes d'une colonne est proportionnelle à la moyenne des taux de mortalité pour une cause fixée, sur l'ensemble des pays considérés. Dans ces conditions la distance du Khi-deux, utilisée par l'analyse factorielle des correspondances est tout à fait adaptée pour étudier les ressemblances entre les répartitions des décès d'un pays à l'autre.

Nous avons donc deux solutions pour le calcul des distances. La première consiste à calculer la distance du Khi-deux directement sur le tableau des données brutes (*Cf.* tableau 1) ; la seconde est de calculer la distance euclidienne usuelle sur les premiers axes issus de l'analyse des correspondances (tableau 2). Dans cette dernière stratégie se pose le problème du nombre d'axes à retenir. Si l'on conserve tous les facteurs possibles (nombre de variables moins un) alors les résultats sont rigoureusement identiques à ceux de la première méthode. Pour apprécier l'effet de "filtrage" de l'analyse factorielle nous préférons ne retenir que trois axes, qui représentent 93.7% de l'inertie totale, le quatrième axe tombant à 4.4% de l'inertie totale.

Les résultats de ces deux séries de calculs figurent dans les tableaux 1 et 2. Etant donnée l'approximation adoptée dans la deuxième méthode, ces deux tableaux ne sont pas facilement comparables si ce n'est en observant l'ordre dans lequel se présentent les distances. Ainsi, en commençant par les plus petites d'entre elles, on a dans le premier cas (distance du Khi-deux sur données brutes) :

```
d(WGERMA,AUSTR) < d(NETHER,ENGLAND) < d(NORW,SCOTL) < d(ICELAN,NORW)
      117               128              152             159
```

Dans le deuxième cas (distance euclidienne sur trois facteurs) :

```
d(WGERMA,AUSTR) < d(NETHER,ENGLAND) < d(ICELAN,NORW) < d(NORW,SCOTL)
      54               67                119             145
```

L'ordre des distances est approximativement le même.

	AUST	FRAN	PORT	WGER	BELG	FINL	SWED	SWIT	ITAL	NIRE	DENM	ICEL
FRANCE	361											
PORTUG	388	440										
WGERMA	117	322	412									
BELGIU	322	347	510	338								
FINLAN	570	638	882	565	417							
SWEDEN	430	384	702	395	268	274						
SWITZE	319	501	670	315	304	295	265					
ITALY	438	453	222	456	630	968	770	749				
NIRELA	1179	1184	1196	1208	1084	1079	1134	1184	1287			
DENMAR	444	604	769	406	422	341	342	176	858	1267		
ICELAN	717	664	909	745	443	435	451	574	1006	1094	675	
SCOTLA	565	472	730	588	308	418	339	495	815	982	620	227
SPAIN	420	302	329	428	548	872	652	689	212	1260	812	874
NORWAY	610	538	829	627	363	356	318	474	904	1088	586	159
SIRELA	684	646	804	745	473	613	581	663	884	1044	808	288
NETHER	464	513	643	498	179	387	349	370	779	1048	476	332
ENGLAN	486	520	702	521	229	315	313	364	813	999	486	266
USA	658	737	730	683	663	717	720	713	806	560	805	857

	SCOT	SPAI	NORW	SIRE	NETH	ENGL
SPAIN	680					
NORWAY	152	762				
SIRELA	280	761	327			
NETHER	275	702	313	392		
ENGLAN	198	717	224	344	128	
USA	687	800	791	809	694	656

Tableau 1. *Données PSYSOC, distances du Khi-2 sur données brutes (multipliées par 1000).*

	AUST	FRAN	PORT	WGER	BELG	FINL	SWED	SWIT	ITAL	NIRE	DENM	ICEL	SCOT	SPAI
FRANCE	218													
PORTUG	339	303												
WGERMA	54	263	370											
BELGIU	277	237	506	312										
FINLAN	557	614	855	564	384									
SWEDEN	355	381	645	369	164	244								
SWITZE	317	433	649	308	274	282	167							
ITALY	433	395	121	455	614	966	750	748						
NIRELA	1170	1173	1184	1207	1079	1077	1128	1180	1283					
DENMAR	422	564	761	398	419	321	296	146	852	1265				
ICELAN	712	623	900	743	435	425	412	572	1004	1091	674			
SCOTLA	546	449	703	586	277	416	324	484	811	982	613	211		
SPAIN	379	263	144	413	494	869	641	670	171	1253	789	857	672	
NORWAY	594	516	796	625	319	355	298	462	901	1085	572	119	145	759
SIRELA	680	536	775	724	423	583	490	653	873	1021	784	244	179	720
NETHER	422	375	642	454	151	304	166	323	752	1030	455	296	161	626
ENGLAN	478	437	695	509	214	281	205	357	807	991	480	260	142	683
USA	652	710	700	682	644	717	703	711	805	553	797	853	684	794

	NORW	SIRE	NETH	ENGL
SIRELA	253			
NETHER	183	332		
ENGLAN	159	320	67	
USA	789	793	656	644

Tableau 2. *Distances euclidiennes usuelles sur les 3 premiers facteurs de l'Analyse factorielle des correspondances (multipliées par 1000)*

2.2.- Phytosociologie (PHYTOS)

Pour l'exemple des données phytosociologiques, on prend l'indice de distance de Jaccard. On aurait pu, également, calculer les distances au sens du Khi-deux. Mais l'expérience montre que les disparités de poids entre espèces provoquent des fluctuations disproportionnées dans les distances et les classifications ultérieures s'en trouvent souvent difficiles à interpréter (*Cf.* chapitre 4, paragraphe 2). Les résultats sont consignés dans le tableau 3, où les valeurs sont multipliées par mille.

	R3	R4	R10	R13	R14	R15	R16	R23	R24	R27	R30	R31	R36	R38	R54
R4	550														
R10	632	629													
R13	590	622	784												
R14	486	514	563	600											
R15	500	675	763	276	543										
R16	474	579	629	658	424	528									
R23	727	732	821	469	718	455	667								
R24	675	744	697	757	531	667	469	531							
R27	721	756	676	800	711	810	725	833	789						
R30	750	789	750	735	625	784	686	811	758	697					
R31	756	763	794	811	636	850	730	902	765	625	500				
R36	537	600	718	675	541	690	634	826	756	528	667	676			
R38	825	868	844	919	861	923	838	975	914	621	654	615	784		
R54	651	579	629	543	595	528	683	632	744	725	789	763	634	868	
R55	585	579	545	692	556	707	615	732	676	622	757	763	564	806	412

Tableau 3. *Données PHYTOS, indices de distance de Jaccard entre relevés (multipliés par 1000)*

3.- Les procédures de calcul de distances

Trois procédures séparées sont proposées dans le classeur Excel : la procédure *DisEuc* pour le calcul des distances euclidiennes usuelles, la procédure *DisKi2* pour le calcul des distances du Khi-2 et la procédure *DisJac* pour le calcul des indices de distance de Jaccard.

La procédure *DisEuc* calcule les distances sur les données telles qu'elles sont présentées dans la feuille active du classeur Excel ; il appartient à l'utilisateur d'effectuer une standardisation préalable des données si cette opération est nécessaire.

En général, dans les trois procédures, les distances sont calculées entre les lignes du tableau. Pour effectuer le calcul entre les colonnes il faut donc recopier les données avec transposition dans une nouvelle feuille. Cependant, la procédure *DisJac* peut calculer les distances de Jaccard sur les lignes ou sur les colonnes. En effet cette procédure est destinée à traiter des données phytosociologiques dans lesquelles il y a souvent un très grand nombre d'espèces. Or si ce nombre dépasse 255 le tableau ne peut pas être disposé avec les espèces en colonnes. Dans cette éventualité on peut mettre les espèces en lignes et les relevés en colonnes (selon l'usage) et travailler tout de même sur les relevés.

Pour la commodité de la lecture et par souci d'homogénéité les résultats se présentent sous la forme d'un tableau carré, symétrique par rapport à la première diagonale, qui, elle, ne comporte que des zéros.

Chapitre 4

La construction ascendante hiérarchique

1.- Généralités

1.1 . - Principe général des constructions ascendantes

On suppose que les distances entre tous les objets, deux à deux, ont été calculées suivant 1'une des formules du chapitre précédent. On procède alors par étapes successives, chacune d'elles consistant à réunir les deux objets les plus. proches. A la fin de chaque étape on recalcule les distances entre le groupe nouvellement créé et le reste des objets. Cela permet de réitérer le processus jusqu'à ce que tous les objets aient été réunis dans un seul groupe. Lorsque cela est achevé on dresse un arbre hiérarchique dont les nœuds représentent les fusions successives, la hauteur de ces nœuds étant égale à la valeur de la distance entre les deux objets, ou groupes, fusionnés. Le niveau des nœuds a donc ainsi une signification concrète ; on dit dans ce cas qu'on obtient une hiérarhie indicée.

La seule difficulté de ce processus reside dans le choix d'une formule pour le recalcul des distances après fusion. Curieusement les considérations mathématiques ne sont pas d'un grand secours pour faire ce choix (voir cependant ci-dessous paragraphe 1.2 et annexe 2). Dans les méthodes usuelles il est plutôt le fruit du bon sens ... et de l'expérience. Nous allons examiner les trois formules les plus courantes. On désigne par i et i' les deux objets, ou groupes d'objets, que l'on veut fusionner et par k un autre point de 1' ensemble :

```
          d(iUi', k) = Min (d(i, k), d(i', k))            (1)
          d(iUi', k) = Max (d(i, k), d(i', k))            (2)
d(iUi', k) = [p(i) d(i,k) + p(i') d(i' ,k)] / [p(i) + p(i')]      (3)
```

La formule (1) indique que la nouvelle distance entre le groupe (i, i'), désigné par iUi', et le point k sera égale à la plus petite des deux distances de i à k et de i' à k. La formule (2) stipule, au contraire, que la nouvelle distance doit être égale à la plus grande des deux anciennes. Enfin la formule (3) dit que la nouvelle distance vaudra la moyenne des distances antérieures. Dans cette formule p(i) et p(i') désignent le nombre d'objets appartenant au groupe i et au groupe i'. Au début de l'algorithme ces groupes sont réduits à un seul point mais il n'en est pas de même au bout de quelques étapes. Ces pondérations assurent qu'à tout moment la distance calculée entre deux groupes est égale à la moyenne des distances initiales entre les points de l'un et les points de l'autre (distances inter-groupes).

D'ailleurs, si l'on n'utilisait pas ces pondérations, on s'exposerait à des désagréments. En effet à chaque étape de l'algorithme on prend la valeur de la distance entre les deux points fusionnés pour niveau du nœud de l'arbre hiérarchique. Les distances recalculées par l'une ou l'autre des formules ci- dessus sont donc des valeurs possibles pour le niveau des nœuds suivants de la hiérarchie. Mais pour que celle-ci puisse être construite il faut que ces niveaux ultérieurs soient supérieurs à celui que l'on vient de créer. On aurait autrement un phénomène "d'inversion" (voir figure 1).

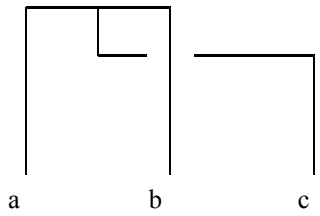


Figure 1.- *Phénomène d'inversion. La distance entre l'élément c et le groupe (a, b) est plus faible que la distance entre a et b.*

Quelle que soit la formule adoptée il faut donc s'assurer que les distances reca1culées soient supérieures au niveau du nœud que 1'on vient de former :

$$d(iUi', k) \geq d(i, i')$$

Cela est évident pour les formules (1) et (2) puisqu'au moment de la fusion d(i, i') est la plus petite de toutes les distances. On vérifie aisément que c'est encore vrai de la formule (3), pour la même raison.

Lorsqu'on utilise la formule (1) , on dit qu'on procède à l'agrégation par "le saut minimum" ou "du lien simple" (en anglais : "single link"), parce que la fusion de deux groupes est basée sur la plus petite des distances inter-groupes. La hiérarchie basée sur la formule (2) est appelée hiérarchie du "diamètre" ou du "lien complet" –(en anglais « complete link »), car elle est basée sur la plus grande distance interne au groupe résultant, ce qui est la définition même du diamètre de ce groupe. Enfin la classification fondée sur la formule (3) s'appelle hiérarchie de « la distance moyenne » (« average link » en anglais).

Deux remarques s'imposent à propos de cette construction hiérarchique :

- les nombreuses recherches et modifications sur les distances obligent à gérer celles-ci en mémoire centrale de l'ordinateur ; ce qui limite sérieusement la taille de l'échantillon.
- en revanche ce type d'algorithme est peu exigeant sur les propriétés de la distance initiale qui peut être obtenue par des formules spéciales (cf. annexe 1) ne satisfaisant pas forcément aux axiomes usuels des distances.

1.2.- Propriétés des formules élémentaires de recalcul

Propriété 1 : Transformation monotone des distances initiales

Soit d(i, i') la distance initiale entre les objets i et i'. Une transformation monotone de ces distances est une modification de d, que nous appellerons d', qui conserve l'ordre entre les distances. C'est à dire que

$$d(i, i') \leq d(j, j') \Rightarrow d'(i, i') \leq d'(j, j')$$

En particulier toute fonction croissante de d a cette propriété. Si l'on applique une telle transformation aux distances initiales, il est clair que l'arbre hiérarchique va être modifié. Cependant dans le cas de l'agrégation par le diamètre ou par le saut minimum, les nœuds successifs vont regrouper les mêmes objets tout au long de l'algorithme. Autrement dit les niveaux de regroupement changent mais la structure de l'arbre hiérarchique est invariante. Ceci relativise la question du choix de l'indice de distance (cf annexe 1). Cette propriété n'est pas vraie pour l'agrégation par la distance moyenne.

Propriété 2 : Extrémalité de la hiérarchie du saut minimum

Lorsqu'on a construit une hiérarchie par l'un des trois procédés ci-dessus on peut en déduire une nouvelle évaluation d* de la distance entre deux objets i et i'. On décide pour cela, que la distance d*(i, i') entre les objets est égale à la hauteur du nœud le plus bas qui assure la liaison entre ces deux objets. On vérifie facilement que les valeurs ainsi établies satisfont aux axiomes mathématiques des distances (voir annexe 1) et en particulier à l'inégalité du triangle. Elles satisfont, en outre, à l'inégalité ultramétrique qui est plus contraignante que celle du triangle :

$$d^*(i, i'') \leq \text{Max}\ (d^*(i, i'), d^*(i', i''))$$

C'est pourquoi on appelle "distance ultramétrique" ou, en abrégé, "ultramètrique", une distance satisfaisant à cette inégalité. On montre (*Cf.* annexe 2) qu'à toute hiérarchie indicée correspond une distance ultramétrique et une seule.

Supposons maintenant que l'on ait deux distances d et d' sur le même ensemble I d'objets. On dit que d est inférieure à d' si, et seulement si, pour tout couple d'objets i et i' on a :

$$d(i, i') \leq d'(i, i')$$

Il est clair que, par construction, l'ultramétrique d, associée à la hiérarchie du saut minimum, est inférieure à la distance d donnée. Mais elle possède en outre l'importante propriété suivante : parmi toutes les ultramétriques inférieures à la distance d, d* est supérieure à toutes les autres. Autrement dit d* s'approche de d "par le bas" le mieux possible (cf annexe 2).

1.3. Comparaison des agrégations par le saut minimum et par le diamètre

Examinons la figure 2 formée de quatre points x, y, z, t, alignés et séparés par des distances voisines : d(x, y) = 1 ; d(x, z) = 2.1 ; d(x, t) = 3.3 ; d(y, z) = 1.1 ; d(y, t) = 2.3 ; d(z, t) = 1.2.

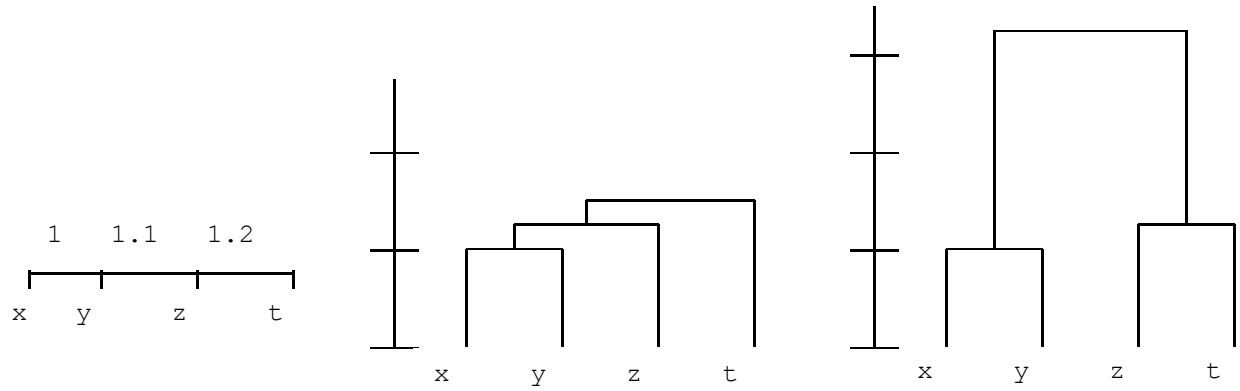


Figure 2. *– Pour les mêmes données où les points sont disposés "en chaîne " (à gauche), les CAH du saut minimum (au centre) et du diamètre (à droite) donnent des résultats radicalement différents.*

Le premier groupe formé est toujours xUy à la distance 1 . Dans l'agrégation par le saut minimum on a :

$$d(xUy, z) = 1.1 \ ; \ d(xUy, t) = 2.3$$

tandis qu'avec l'agrégation par le diamètre :

$$d(xUy, z) = 2.1 \ ; \ d(xUy, t) = 3.3$$

Dans le premier cas on agrège z à xUy, tandis que dans le second on agrège z et t distants seulement de 1.2 . La dernière étape consiste à réunir tous les objets, d'où les graphiques ci-dessus. On remarque que l'agrégation par le saut minimum a tendance à " écraser " les niveaux de liaison, tandis que la méthode du diamètre les distend. Avec le saut minimum on conçoit que 1'on arrive à rapprocher des points extrêmement différents ; c'est ce qu'on appelle "l'effet de chaîne"

2. Application aux exemples

2.1.- Causes de décès (PSYSOC)

On a appliqué la construction ascendante sur les deux matrices de distances entre pays calculées au chapitre précédent. La première était obtenue en calculant la distance du Khi-deux sur les données brutes, tandis que la seconde provenait de la formule euclidienne usuelle appliquée aux résultats de 1'AFC. Les deux résultats (voir figures 3 et 4) sont très semblables et font apparaître trois groupes principaux :

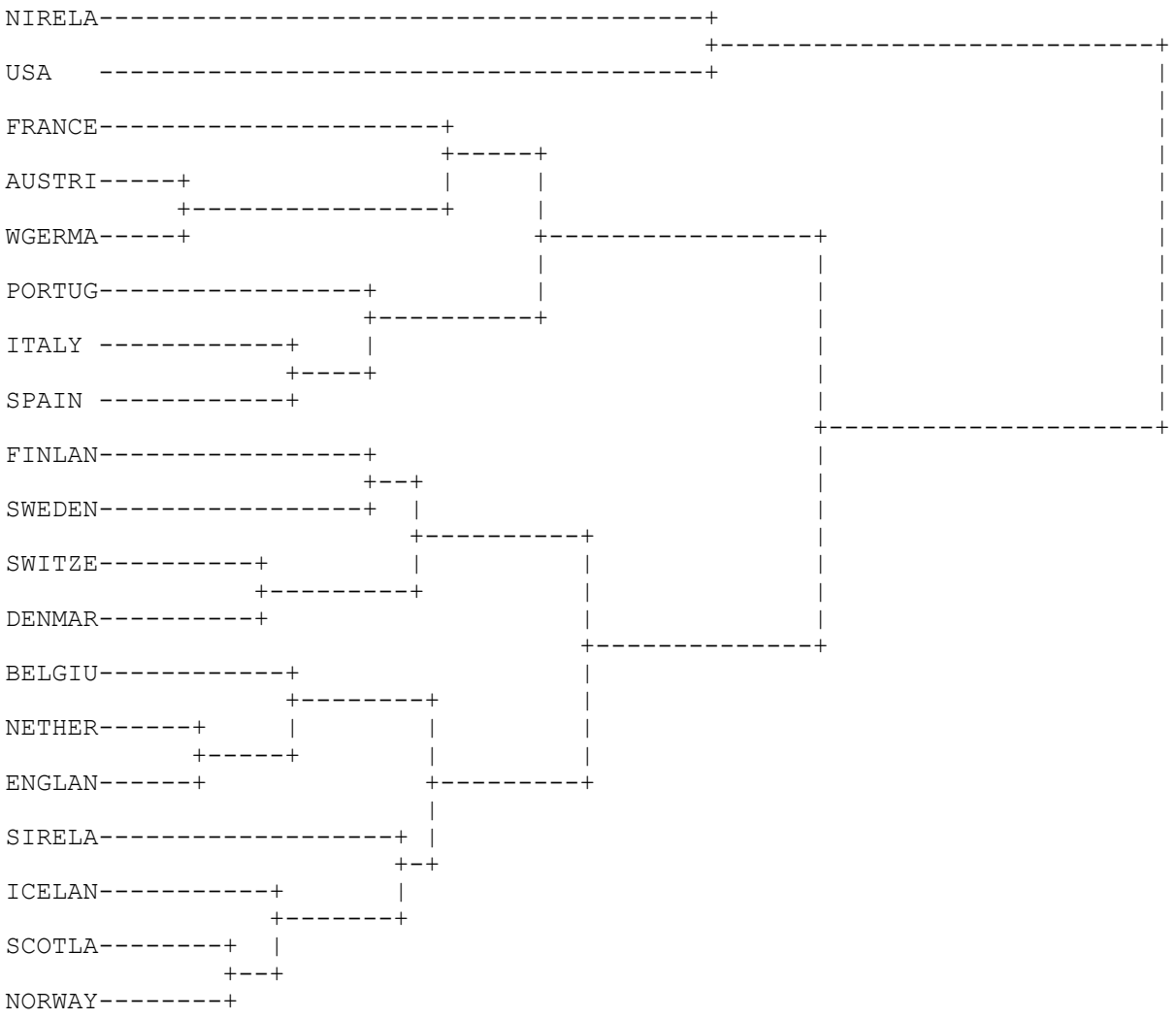


Figure 3. *- Données PSYSOC. Hiérarchie du lien moyen, construite à partir de la distance du Khi-2 calculée sur les données brutes.*

- groupe "Europe-Ouest" : AUSTRI, WGERMA, FRANCE, SPAIN, ITALY, PORTUG
- groupe "Europe-Nord" : ICELAN, NORWAY, NETHER, ENGLAN, SCOTLA, SIRELA, BELGIU, SWEDEN, SWITZE, DENMAR, FINLAN
- groupe "Atlantique" : USA et NIRELA

Les deux premiers groupes sont subdivisés en deux sous-groupes que l'on distingue aisément par l'importance de l'écart entre les niveaux de jonction. Il est intéressant de constater que, dans les deux calculs, la FRANCE ne se rattache pas à ses "soeurs latines" que sont l'ITALY et SPAIN, mais à AUSTRI et WGERMA. Ce qui confirme la thèse de Mr E. Todd , qui soutient, contrairement aux idées reçues, que France et Allemagne se ressemblent beaucoup quant aux comportements sociaux.

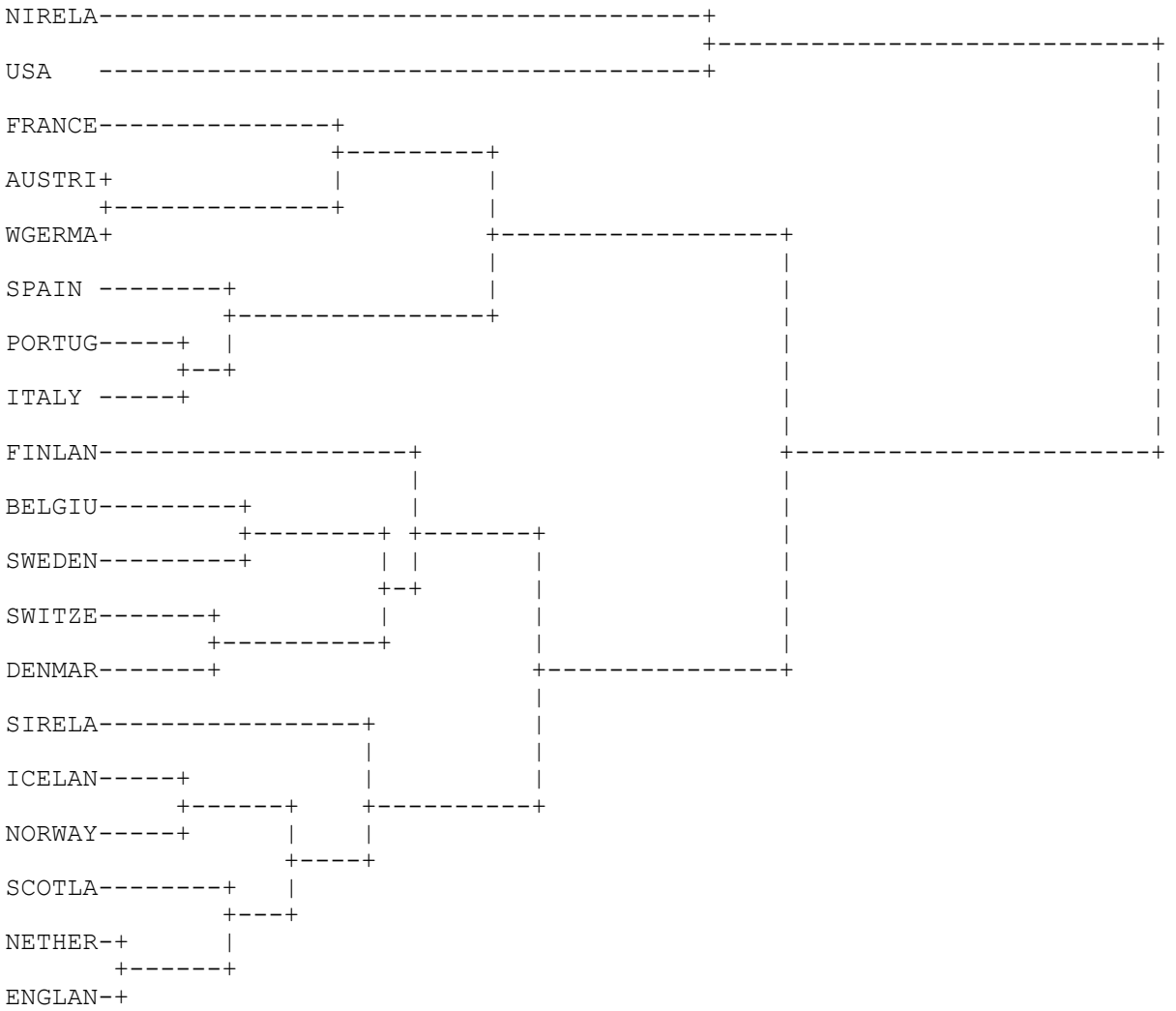


Figure 4. *- Données PSYSOC. Hiérarchie du lien moyen, construite à partir de la distance euclidienne usuelle calculée sur les coordonnées factorielles (A.F.C., 3 facteurs)*

2.2.- Phytosociologie (PHYTOS)

La comparaison des deux arbres hiérarchiques obtenus, l'un en agrégeant par la distance moyenne, l'autre par le diamètre (lien maximum), fait apparaître les mêmes groupes principaux, au nombre de quatre, qui coïncident assez bien avec les groupes établis lors de l'étude antérieure (Voir Chapitre 2, paragraphe 1.2). Sur les deux figures 5 et 6 apparaissent les rassemblements suivants :

3, 4, 14 16 (PAcn1)
10, 54, 55 (PAcn2)
13, 15, 23 (PAn)
27, 30, 31, 36, 38 (PAc)

Le relevé numéro 24 est isolé à l'extrémité d'une longue branche dont le rattachement change selon le mode d'agrégation employé. Il s'agit clairement d'un relevé intermédiaire d'affectation délicate.

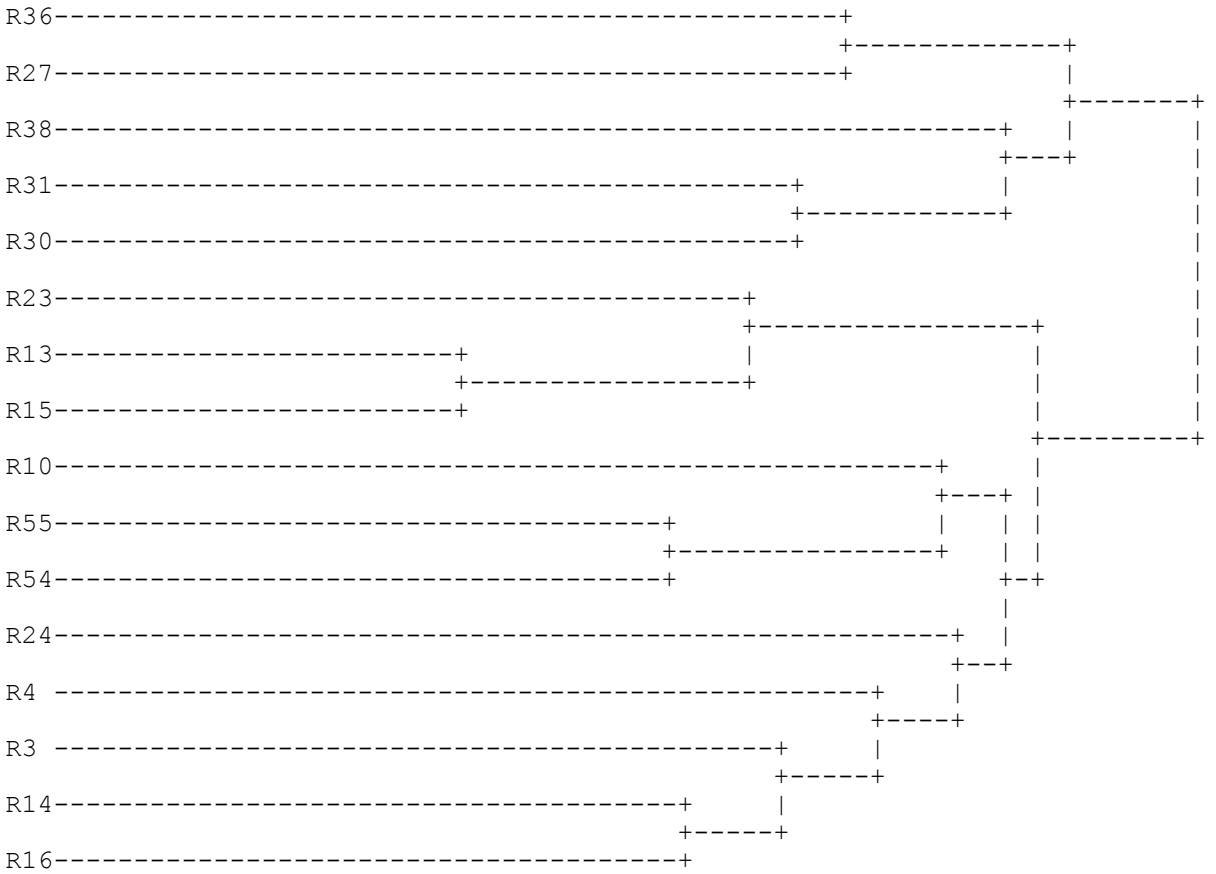


Figure 5.- *Données PHYTOS. Hiérarchie du lien moyen, basée sur la distance de Jaccard.*

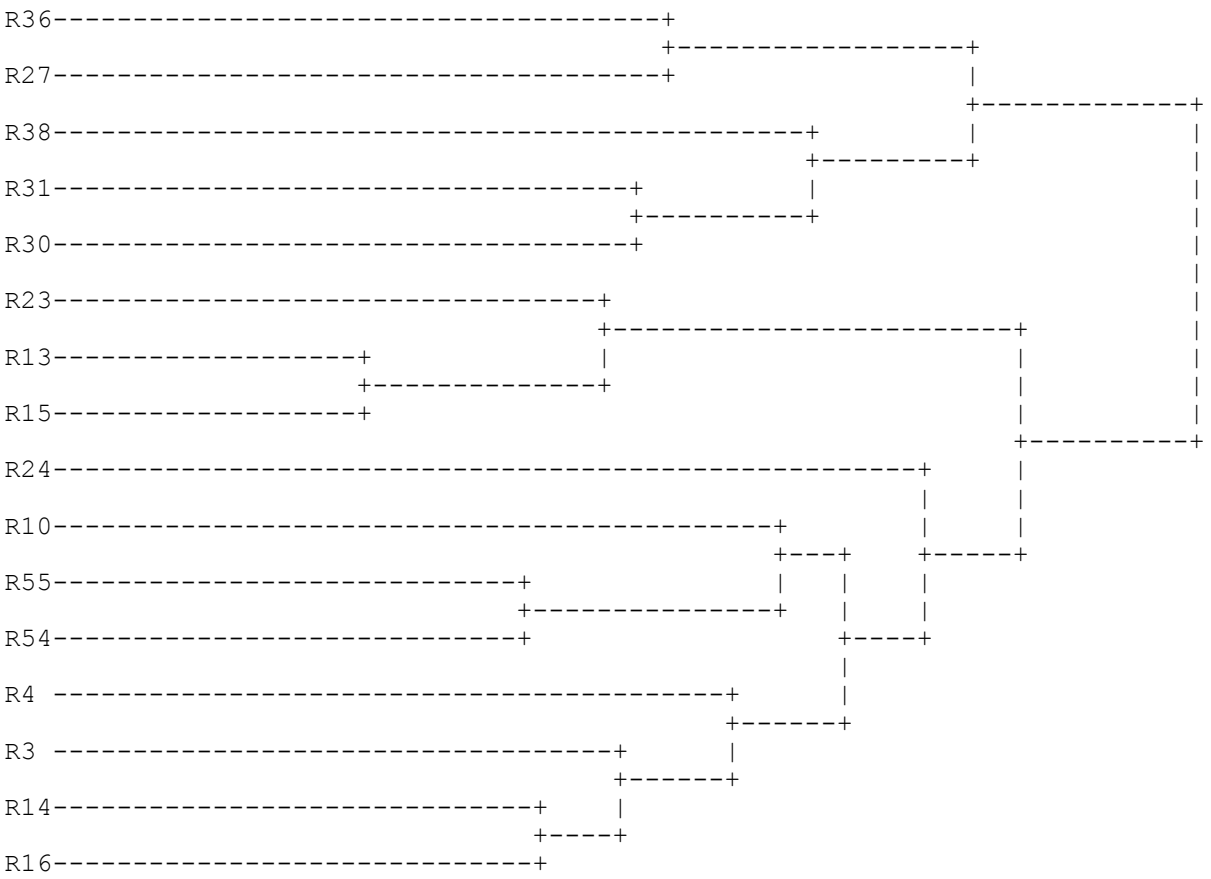


Figure 6.- *Données PHYTOS. Hiérarchie du diamètre, basée sur la distance de Jaccard.*

3.- Les procédures de construction ascendantes de hiérarchies

Les procédures Excel suivantes sont disponibles dans le classeur « AnaDon.xls » .

CAHLM : calcule la CAH du lien moyen
CAHdiam : calcule la CAH du diamètre (ou lien complet)
CAHsmin : calcule la CAH du saut minimum (ou lien simple)
DessArb : dessine l'arbre hiérarchique obtenu par les méthodes précédentes.

Chapitre 5

Agrégation autour de centres mobiles

1.- Principes et problèmes

1.1.- L'algorithme des centres mobiles

L'algorithme que nous allons décrire a pour but de construire une seule partition de l'ensemble étudié. Il en existe de nombreuses variantes mais nous ne parlerons que de la plus simple d'entre elles.

Au début de l'algorithme il faut se fixer un nombre k de classes et choisir une partition initiale. Cette partition peut être inspirée par une connaissance *a priori* des objets à classer ; ou bien elle peut être obtenue par répartition au hasard des objets en k catégories. On exécute alors les opérations suivantes:

1) Pour chaque classe q déterminer le centre de gravité g_q
2) Réaffecter chaque objet i à la classe C(i) dont le centre de gravité est le plus proche

$$C(i) = q \text{ si et seulement si } d(i, g_q) = \min\{d(i, g_r) | \ r \in Q\}$$

3) Retourner en 1 tant que surviennent des modifications dans la composition des classes.

Cet algorithme très simple repose sur d'intéressantes propriétés mathématiques que l'on va examiner maintenant. Ses avantages et inconvénients seront discutés au paragraphe 1.3. Les développements mathématiques inhabituels du paragraphe 1.2 sont nécessaires car ils seront utilisés également au chapitre suivant, qui expose une construction ascendante hiérarchique importante par la qualité des résultats qu'elle fournit.

1.2.- Moment d'ordre deux d'une partition

Par souci de simplification on suppose que toutes les variables, au nombre de p, sont quantitatives et que la dissemblance entre les objets est correctement mesurée par la distance euclidienne d usuelle. On appelle x(i, j) la valeur de la j-ème variable pour la i-ème observation. On suppose, en outre, que ces observations, au nombre de n, sont pondérées par des masses, notées m_i, proportionnelles au rôle que l'on veut leur faire jouer. Par exemple, si l'observation i représente l'individu moyen d'une sous-population, on peut décider que m_i est l'effectif de la sous-population. S'il n'y a pas lieu de pondérer les observations on affectera la valeur 1 à tous les m_i.

De la sorte on peut se représenter les observations comme un nuage matériel I formé des masses ponctuelles m_i. Son centre de gravité g a pour j-ème coordonnée :

$$x(g, j) = [x(1, j) + x(2, j) + ...+ x(n, j)] / m$$

où $m = m_1 + m_2 + \ldots + m_n$ est la masse totale du nuage.

Remarquons, en passant, que x(g, j) n'est autre que la moyenne de la variable j. Le moment centré d'ordre deux, ou moment par rapport au centre de gravité, est la quantité :

$$M^2(I/g) = m_1\, d^2(1, g) + m_2\, d^2(2, g) + \ldots + m_n\, d^2(n, g) \qquad (5.1)$$

où $d^2(i, g)$ désigne le carré de la distance de i à g. Dans le cas de la distance euclidienne usuelle :

$$d^2(i, g) = (x(i, 1) - x(g, 1))^2 + (x(i, 2) - x(g, 2))^2 + \ldots + (x(i, p) - x(g, p))^2$$

Autrement dit, le moment centré d'ordre deux du nuage I s'obtient comme la somme, pour toutes les variables et tous les objets, des carrés des écarts à la moyenne (somme pondérée par les masses des objets). C'est une mesure de la dispersion des points du nuage. En effet, si les points sont très concentrés autour de leur centre de gravité, le moment d'ordre deux est faible, il est grand dans le cas contraire. D'ailleurs la variance d'une variable j, qui est la mesure usuelle de la dispersion en statistique s'écrit :

$$var(j) = [m_1\, (x(1, j) - x(g, j))^2 + \ldots + m_n\, (x(n, j) - x(g, j))^2] / m$$

C'est la moyenne pondérée de la somme des carrés des écarts pour la variable considérée. Au coefficient 1/m près, le moment d'ordre deux est donc une variance généralisée au cas de p variables. Ce moment d'ordre 2 est encore appelé "Moment d'inertie" car il est correspond exactement à cette notion de mécanique.

Théorème de Huyghens

Examinons maintenant le cas du moment d'ordre deux par rapport à un point a, différent du centre de gravité.

$$M^2(I/a) = m_1\, d^2(1, a) + m_2\, d^2(2, a) + \ldots + m_n\, d^2(n, a)$$

Le i-ème terme $m_i\, d^2(i, a)$ de cette somme est, lui-même, une somme pondérée de carrés d'écarts aux coordonnées x(a, j) de a, l'indice j parcourant l'ensemble 1, 2, ..., p, des variables.

$$m_i\, d^2(i, a) = m_i\, (x(i, 1) - x(a, 1))^2 + m_i\, (x(i, 2) - x(a, 2))^2 + \ldots + m_i (x(i, p) - x(a, p))^2$$

Le j-ème terme de cette expression peut à son tour se décomposer en faisant intervenir la j-ème coordonnée du centre de gravité :

$$m_i\, [x(i, j) - x(a, j)]^2 = m_i [x(i, j) - x(g, j) + x(g, j) - x(a, j)]^2.$$

$$m_i\, [x(i, j) - x(a, j)]^2 = m_i [x(i, j) - x(g, j)]^2 + 2m_i\, [x(i, j) - x(g, j)][x(g, j) - x(a, j)] + m_i\, [x(g, j) - x(a, j)]^2$$

Pour obtenir le moment d'ordre deux il faudra donc faire une double somme d'expressions analogues : l'une sur les variables (indice j), l'autre sur les individus (indice i).

Commençons par la somme sur les individus et examinons le terme intermédiaire :

$$2m_i\, [x(i, j) - x(g, j)][x(g, j) - x(a, j)].$$

Comme le deuxième crochet ne dépend pas des individus on pourra le mettre en facteur dans la somme des termes intermédiaires qui devient :

$$2[x(g, j) - x(a, j)]\, [m_1\, (x(1, j) - x(g, j)) + m_2\, (x(2, j) - x(g, j)) + \ldots + m_n\, (x(n, j) - x(g, j))]$$

Mais la deuxième expression entre crochets est nulle de par la définition du centre de gravité (La somme des écarts à la moyenne est égale à zéro). Revenons alors à la double somme constituant le

moment d'ordre deux. Les deux types de termes restants fournissent, l'un, le moment centré d'ordre deux, l'autre, une quantité qui s'écrit m $d^2(g, a)$

$$M^2(I/a) = M^2(I/g) + m\ d^2(g, a) \qquad (5.2)$$

C'est le théorème de Huyghens qui s'énonce ainsi : le moment d'inertie d'un solide, par rapport à un point quelconque a, est égal au moment du solide par rapport à son centre de gravité augmenté du moment du point g, affecté de la masse totale m du solide, par rapport au point a.

Application à une partition

Supposons définie une partition Q de l'ensemble I ; c'est à dire que tout élément q de Q est un sous-ensemble de I, et tout élément de I appartient à un et un seul des éléments de Q. On appelle m_q la masse du sous-ensemble des points de q. Reprenons dans une écriture condensée l'expression 5.1 du moment centré (le signe Σ_i signifie qu'il faut faire la somme de tous les termes analogues obtenus en faisant varier l'indice i) :

$$M^2(I/g) = \Sigma_i\ m_i\ d^2(i, g)$$

et décomposons cette somme en faisant des sommes partielles sur les sous-ensembles q de Q :

$$M^2(I/g) = \Sigma_{q \in Q}\ [\Sigma_{i \in q}\ m_i\ d^2(i, g)]$$

La somme entre crochets représente le moment de la classe q par rapport au point g, centre de gravité général, qui est différent du centre de gravité g_q de cette classe. On peut donc appliquer le théorème de Huyghens pour le sous-ensemble q :

$$M^2(I/g) = \Sigma_q\ [M^2(q/g_q) + m_q\ d^2(g_q, g)]$$

que l'on peut encore écrire :

$$M^2(I/g) = \Sigma_q\ M^2(q/g_q) + M^2(Q/g) \qquad (5.3)$$

En effet, la deuxième somme, issue du crochet, n'est autre que le moment centré d'ordre deux du solide formé par les centres de gravité g_q, chacun d'eux étant muni de la masse m_q, car ce solide a son centre de gravité confondu avec le centre de gravité g de I.

L'équation (5.3) représente la décomposition de la dispersion totale en dispersion à l'intérieur des classes, appelée intra-classe, et dispersion entre les classes, ou inter-classe. On dit que le moment d'ordre deux total est égal à la somme des moments centrés de chacune des classes, augmentée du moment inter-classe. Cette équation est analogue à celle de l'Analyse de la variance dans le cas d'une seule variable.

Il est évident qu'une bonne classification doit rendre la dispersion intra-classe aussi petite que possible, pour fournir des classes homogènes. La dispersion totale étant fixée par les données elles-mêmes, il est équivalent de chercher une partition minimisant la dispersion intra-classe ou rendant maximum la dispersion inter-classe. L'une ou l'autre de ces quantités constitue le critère du moment d'ordre deux d'une partition. On en verra une application à la construction ascendante hiérarchique dans le chapitre 6.

Application à l'algorithme des centres mobiles.

Examinons ce que devient le moment intra-classe W au cours du déroulement de l'algorithme. Dans la phase de réaffectation des objets, appelons q* la classe reconstituée autour du centre de gravité g_q de l'ancienne classe q.

$$W = \Sigma_{q \in Q} \quad \Sigma_{i \in q} \; m_i \; d^2(i, \; g_q)$$

Appelons S la valeur de W après réaffectation des points i au centre de gravité le plus proche :

$$S = {}_{\Sigma q \in Q \quad \Sigma i \in q^* \; mi} \; d^2(i, \; g_q)$$

Soit i un élément de la classe q. Si i n'a pas changé de classe, sa contribution au moment intra-classe reste la même. Mais s'il provient d'une autre classe q' alors c'est qu'il est plus proche de g_q que de $g_{q'}$ donc $d^2(i, g_q) < d^2(i, g_{q'})$ et sa contribution à S est inférieure à celle qu'il avait dans W. Il en résulte que S < W. Remarquons que S n'est plus la somme des moments centrés puisque les g_q ne sont plus les centres de gravité des classes q^*. Considérons alors la valeur W* du moment intra-classe de la nouvelle partition :

$$W^* = \Sigma_{q \in Q} \quad \Sigma_{i \in q^*} \; m_i \quad d^2(i, \; g_{q^*})$$

Cette fois on prend en compte les moments centrés qui sont, d'après le théorème de Huyghens, inférieurs aux moments non centrés. Donc W* < S. Il en résulte qu'à la fin de cette étape le moment intra-classe W* est inférieur à ce qu'il était à la fin de l'étape précédente et la nouvelle partition est donc meilleure que la partition précédente.

Cela ne veut pas dire pour autant que la partition finale de l'algorithme des centres mobiles soit la meilleure partition possible en k classes. En effet, la partition finale dépend de la partition initiale. Une autre partition initiale peut donc donner une partition finale pour laquelle le critère du moment d'ordre deux soit encore meilleur. On résume cela en disant qu'on obtient un optimum local du critère et non un optimum absolu.

1.3.- Avantages et inconvénients de la méthode

L'algorithme des centres mobiles, contrairement à de nombreuses méthodes classificatoires, a l'avantage d'optimiser un critère simple de dispersion, savoir le moment d'ordre deux d'une partition. Cependant, comme on vient de le voir, on n'a pas la certitude d'obtenir un optimum absolu, c'est à dire la meilleure solution. L'un des moyens généralement préconisés (*cf* Diday, 1971) pour obtenir des résultats valables est d'exécuter plusieurs fois l'algorithme complet, avec des partitions initiales différentes. On peut alors retenir la partition finale associée au moment intra-classe le plus petit (qui n'est pas pour autant le minimum absolu, ce que l'on ne sait pas déterminer).

Cependant une autre stratégie est de procéder à l'examen des "formes fortes". Celles-ci sont constituées des sous-ensembles d'objets qui ont toujours été réunis dans la même classe finale au cours des différents essais de partitions initiales. Ces formes fortes représentent donc des groupes homogènes et mettent en relief les objets d'attribution indécise qui n'appartiennent à aucune forme forte. L'étude des formes fortes permet également de s'affranchir d'un autre inconvénient de la méthode qui est de nécessiter le choix *a priori* du nombre de classes. En effet le nombre de formes fortes peut être très variable et ne dépend pas directement du nombre de classes choisi.

Un autre problème est celui du choix d'une partition initiale. Il est évident que si l'on a des informations sur les regroupements possibles il vaut mieux en tenir compte pour démarrer avec une bonne partition. Notons à ce propos qu'il n'est pas nécessaire d'affecter tous les objets à une classe. On peut laisser certains objets sans affectation. A la première étape de l'algorithme les centres de

gravité seront calculés sur les seuls objets appartenant à une classe déclarée. Puis l'ensemble des objets sera affecté ou réaffecté en fonction de ces centres de gravité.

Signalons enfin une variante possible de cet algorithme. Pour chaque classe trouvée au cours d'une étape on peut prendre un certain nombre, fixé à l'avance, de représentants de cette classe, au lieu du centre de gravité. On réaffecte ensuite l'ensemble des objets en fonction de la moyenne de leurs distances à ces représentants. Les représentants sont des points "centraux", choisis suivant le même critère de la moyenne des distances. Cette variante a l'avantage d'éviter de fabriquer des classes "creuses" ; le centre de gravité peut en effet tomber dans une zone de faible densité, intermédiaire entre deux régions denses.

2.- Application à l'exemple PSYSOC

Dans cette application, plutôt que le moment intraclasse, nous utilisons, comme critère de qualité de la partition obtenue, le rapport R du moment interclasse au moment total. Ce rapport, que nous appellerons "rapport d'inerties de la partition", est toujours compris entre zéro et 1, puisque le moment total s'écrit comme la somme des moments interclasse et intraclasse. Une bonne partition sera donc caractérisée par une valeur de R proche de 1.

Le premier choix délicat de l'algorithme des centres mobiles est celui du nombre de classes de la partition. La construction ascendante hiérarchique nous a permis de déceler (chapitre 4, paragraphe 2) l'existence de trois groupes que l'on a dénommés Méditerranée, Europe-Nord et Atlantique par commodité. Nous avons donc fait une première série de calculs en fixant le nombre de classes à trois, puis une autre série avec quatre classes, pour examiner le comportement du programme dans une situation embarrassante. Dans tous les cas les données sont constituées des trois premiers facteurs de l'Analyse des correspondances.

2.1.- Partition en trois classes

En introduisant, comme partition initiale, les trois groupes déterminés par la construction hiérarchique, le programme ne fait qu'une seule étape qui montre que cette partition initiale ne peut pas être améliorée. Nous avons alors tiré au hasard quatre partitions initiales différentes ; trois d'entre elles ont convergé vers la même partition finale déjà trouvée. Une seule d'entre elles a donné une partition différente (voir tableau 1), mais avec un rapport moment interclasse / moment total de 0.51, beaucoup plus faible que celui de 0.70 qui correspond à la partition précédente. Ce rapport R nous permet de trancher en faveur de la partition trouvée à l'aide de la CAH.

2.2.- Partition en quatre classes

Nous avons voulu voir quels résultats on obtient lorsque l'on choisit un nombre de classes en désaccord avec les données. Ce qui peut arriver, en pratique, si l'on n'a fait aucune analyse préalable. On a effectué quatre tirages au hasard en quatre classes. Les partitions P1, P2, P3, P4 issues de ces tirages, ainsi que les rèsultats P1*, P2*, P3*, P4*, de l'algorithme des centres mobiles, sont consignés dans le tableau 2. Les rapports d'inertie obtenus sont respectivement : 0.80, 0.49, 0.75 et 0.81. La partition P4* est donc la meilleure, mais elle est suivie de près par P1* et P3*. La partition P2* est franchement mauvaise.

	P	P*
AUSTRI	1	1
FRANCE	3	1
PORTUG	3	1
WGERMA	1	1
BELGIU	1	3
FINLAN	3	3
SWEDEN	3	3
SWITZE	3	3
ITALY	1	1
NIRELA	2	2
DENMAR	2	3
ICELAN	1	2
SCOTLA	1	2
SPAIN	1	1
NORWAY	2	2
SIRELA	3	2
NETHER	2	3
ENGLAN	2	2
USA	1	2

	P1	P3	P4	P1*	P3*	P4*	P2	P2*
AUSTRI	4	1	1	1	1	1	1	2
FRANCE	1	4	1	1	1	1	2	3
PORTUG	1	4	1	1	2	1	4	1
WGERMA	4	1	1	1	1	1	1	2
BELGIU	3	1	3	3	4	2	4	3
FINLAN	3	4	3	3	4	2	3	4
SWEDEN	4	4	4	3	4	2	4	3
SWITZE	3	4	2	3	4	2	2	2
ITALY	4	2	1	1	2	1	1	1
NIRELA	4	3	3	2	3	3	1	1
DENMAR	1	2	2	3	4	2	4	2
ICELAN	4	1	2	4	4	4	1	4
SCOTLA	4	1	2	4	4	4	4	4
SPAIN	4	1	2	1	2	1	3	1
NORWAY	3	3	1	4	4	4	4	4
SIRELA	2	4	3	4	4	4	4	4
NETHER	2	3	4	3	4	4	2	4
ENGLAN	4	3	2	3	4	4	3	4
USA	2	1	2	2	3	3	4	1

Tableau 1 (à gauche) et tableau 2 (à droite).

Tableau 1 : Partitions initiale (P) et finale (P) en trois classes. R = 0.51*

Tableau 2 : Partitions initiales (P1, P2, P3 et P4) et finales (P1, P2*, P3* et P4*) en quatre classes. R1 = 0.8 ; R2 = 0.49 ; R3 = 0.75 ; R4 = 0.81. La partie encadrée en gras correspond aux trois meilleures partitions finales, sur lesquelles sont basées les formes fortes énumérées dans le tableau 3 ci-dessous.*

L'examen attentif des trois meilleures partitions montre que celles-ci ressemblent beaucoup à la "bonne" partition en trois classes obtenue précédemment. Elles s'obtiennent par scission de l'une des trois classes. Ainsi P1* coupe en deux le groupe "Europe-Nord", tandis que P3* subdivise le groupe "Méditerranée", enfin P4* scinde encore le groupe "Europe-Nord" mais d'une manière diffèrente de P1*.

Il est facile de déterminer, à la main, les groupements stables ou formes fortes, en repérant les pays ayant la même succession de numéros de classe à travers les trois partitions retenues (voir tableau 3). Ceci nous conduit à six groupements, qui ne sont pas en contradiction avec les hiérarchies déjà obtenues. Aucun pays ne reste isolé. Il est malheureusement impossible de dire si une partition en

six classes est meilleure qu'une autre en trois classes, car le rapport d'inerties, qui nous sert de critère, dépend du nombre de classes, comme le font d'autres critères non basés sur l'inertie.

G1 = (1, 1, 1) : AUSTRI, FRANCE, WGERMA
G2 = (1, 2, 1) : PORTUG, ITALY, SPAIN
G3 = (3, 4, 2) : BELGIU, FINLAN, SWEDEN, SUISS, DENMAR
G4 = (2, 3, 3) : NIRELA, USA
G5 = (4, 4, 4) : ICELAN, SCOTLA, SIRELA, NORWAY
G6 = (3, 4, 4) : NETHER, ENGLAN

Tableau 3.- *Groupements stables (formes fortes) après tirages de partitions aléatoires en 4 classes. Les pays rassemblés dans un même groupe se sont toujours trouvés ensemble dans les trois partitions finales retenues à l'issue des différents tirages initiaux ; ces groupes sont symbolisés par leurs numéros figurant entre parenthèses (Voir Tableau 2 ci-dessus).*

3.- Les programmes de calculs de Centres mobiles

Nous proposons, dans notre bibliothèque de procédures (Classeur «AnaDon.xls»), deux versions de l'agrégation autour de centres mobiles dénommées respectivement *CenMob1* et *CenMob2*. Dans *CenMob1* l'utilisateur doit fournir une partition initiale qui est alors améliorée par le programme, tandis que dans *CenMob2* l'utilisateur fournit seulement le nombre de classes désiré. Le programme effectue alors un certain nombre (fixé par l'utilisateur) de tirages aléatoires de partitions initiales qui sont toutes soumises à l'algorithme des centres mobiles. Seule la meilleure partition finale est conservée et affichée.

Dans les deux versions le résultat rend compte de la qualité de la partition obtenue, en donnant le rapport d'inerties. Il indique également la contribution de chaque classe au moment inter-classe.

Chapitre 6

Construction ascendante hiérarchique du moment d'ordre deux

1.- Principe et problèmes

La construction hiérarchique du moment d'ordre deux est une méthode agrégative analogue à celles qui sont décrites au chapitre 4. Elle est connue dans le monde anglo-saxon sous le nom de méthode de Ward (Ward, 1963). Son originalité provient de ce que le critère permettant de décider de la fusion de deux classes n'est pas basé sur une quelconque notion de distance entre classes mais sur l'augmentation de la dispersion intra-classe. Pour comprendre cela il nous faut reprendre le théorème de Huyghens, examiné au chapitre précédent (paragraphe 1.2) et l'appliquer au cas particulier d'une partition en deux classes q et q'. Dans ce cas la formule 5.3 du chapitre 5 devient :

$$M^2(qUq') = M^2(q) + M^2(q') + m_q\, d^2(g, g_q) + m_{q'}\, d^2(g, g_{q'})$$

où l'on désigne par qUq' la réunion des deux classes q et q'. On montre par ailleurs facilement (*cf* Benzécri 1975, Jambu 1978), que le moment intra-classe, représenté par les deux derniers termes de la somme ci-dessus, s'écrit aussi :

$$m_q\, d^2(g, g_q) + m_{q'}\, d^2(g, g_{q'}) = [(m_q\, m_{q'})/(m_q + m_{q'})]\, d^2(g_q, g_{q'}) \qquad (6.1)$$

Si les deux classes q et q' sont les éléments d'une partition, cette expression représente l'augmentation du moment intra-classe qui arriverait si l'on fusionnait les deux classes q et q' ; en effet lorsque q et q' sont séparées leur contribution au moment intra-classe vaut $M^2(q) + M^2(q')$

C'est précisément cette quantité (6.1) que l'on prend comme critère d'agrégation dans la hiérarchie du moment d'ordre deux. A chaque pas de l'algorithme on fusionne les deux classes qui provoquent la plus faible augmentation du moment intra-classe. Cette augmentation du moment intra-classe joue donc maintenant le rôle de la distance dans l'algorithme élémentaire (du chapitre 4), nous l'appellerons pseudo-distance. Au début de l'algorithme, supposant que chaque objet est muni d'une masse unité, la matrice des pseudo-distances vaut pour la case (i, i') :

$$(1/2)\, d^2(i, i')$$

Au premier pas de l'algorithme on agrège la paire pour laquelle cette quantité est la plus petite, qui, en l'occurence, coïncide avec celle de la plus petite distance. Pour pouvoir procéder à l'agrégation suivante il faut alors calculer l'augmentation du moment intra-classe avec chacun des autres objets. La formule (6.1) fait intervenir les centres de gravité des classes et ne permet donc pas facilement le recalcul des nouvelles pseudo-distances à partir des anciennes. Il existe heureusement une formule, un peu plus compliquée, qui permet de faire cette mise à jour, donc de suivre, de très près, l'algorithme élémentaire décrit au chapitre 4 :

$$d(iUi',k) = (1/m)\, [(m_i + m_k)\, d(i, k) + (m_{i'} + m_k)\, d(i', k) - m_k\, d(i, i')] \qquad (6.2)$$

m est mis pour la somme $(m_i + m_{i'} + m_k)$ des effectifs des trois groupes en présence. L'écriture d (iUi', k) désigne maintenant la pseudo-distance, c'est à dire l'accroissement du moment intra-classe, qui résulterait de la fusion éventuelle du groupe (iUi'), que l'on vient de former, avec le groupe k. (Voir Benzécri 1973 pour une démonstration).

Cependant nous n'utiliserons pas cette formule. Nous préférons étudier ici un autre algorithme, fournissant les mêmes résultats, mais travaillant directement sur le tableau des données brutes (à supposer que celles-ci soient quantitatives). Ou, mieux encore, sur le tableau des coordonnées issues d'une analyse factorielle, ainsi qu'on l'a recommandé au chapitre 5. Cet algorithme consiste à tenir en mémoire centrale le tableau rectangulaire des données, puis, au fur et à mesure des agrégations, à remplacer les lignes des objets agrégés par une ligne contenant les coordonnées de leur centre de gravité.

L'avantage de cette méthode est qu'elle permet de traiter des ensembles d'objets beaucoup plus importants que l'algorithme élémentaire. En effet lorsque les objets sont nombreux, le nombre des variables est généralement restreint. Supposons, par exemple, qu'on ait 200 objets repérés par 10 variables quantitatives, alors la matrice des données n'occuppe que 200x10 = 2000 cases , tandis que la matrice des distances utilise (200 x 199)/2 = 19900 cases (en ne conservant que la demi-matrice inférieure ou supérieure) ... Dans le cas où le nombre de variables est, lui aussi, élevé alors il devient indispensable d'effectuer une analyse factorielle préalable dont on ne retient que les cinq ou dix premiers axes factoriels.

En contre-partie le nombre de calculs à effectuer sera nettement plus élevé, puisqu'après chaque agrégation il faudra recalculer les pseudo-distances, non seulement entre la paire fusionnée et les autres objets, mais aussi entre tous les objets, puisqu'on ne garde pas en mémoire cette matrice des pseudo-distances. En fait, on va voir que, grâce à la considération des "voisins réciproques", on peut réduire substantiellement cette quantité de calculs et obtenir un algorithme particulièrement efficace.

2.- L'algorithme des "voisins réciproques" (De Rham 1980)

Le plus proche voisin i' d'un objet i est celui pour lequel la distance d(i, i') est la plus petite des distances entre i et tout autre objet. (On élimine le cas, peu courant, où, par suite de distances égales, un objet i aurait plusieurs plus proches voisins). On appelle "voisins réciproques" deux objets dont l'un est le plus proche voisin de l'autre et vice versa. L'algorithme des voisins réciproques est basé sur la propriété suivante :

Soient i et i' les deux objets ou groupes fusionnés, à une étape quelconque de l'algorithme usuel, et k un troisième objet ou groupe :

$$d(iUi', k) \geq \mathrm{Min}(d(i, k), d(i', k)) \qquad (6.3)$$

Cette écriture revient à dire que la formule de recalcul des distances est telle que toute distance recalculée est plus grande que la plus petite de celles qu'elle remplace.

Cette propriété est vérifiée par les trois formules élémentaires examinées au chapitre 4. Montrons que cela est encore vrai pour la formule (6.2) ci-dessus. En effet, on remarque tout d'abord que, puisque i et i' sont agrégés on a :

$$d(i, i') \leq d(i, k) \text{ et } d(i, i') \leq d(i', k)$$

Donc, en remplaçant d(i, i') par d(i, k) ou par d(i', k) on diminue la valeur de l'expression de droite de (6.2). Supposons maintenant que d(i, k) soit inférieur ou égal à d(i', k), alors en remplaçant d(i', k) par d(i, k) le terme de droite dans (6.2) est rendu encore plus petit que sa vraie valeur ; mais ces deux remplacements rendent ce terme égal à d(i, k) qui est donc inférieur ou égal à d(iUi', k). Il en serait de même si d(i', k) était inférieur à d(i, k).

On montre, que, dans ce cas, deux objets qui sont voisins réciproques constituent nécessairement un nœud de la hiérarchie, quelle que soit la distance qui les sépare. On profite alors de cette observation pour agréger, à chaque étape de l'algoritnme, toutes les paires de voisins réciproques, au lieu de la seule paire qui présente la plus petite distance. On réduit ainsi le nombre d'étapes à accomplir et, surtout, on diminue considérablement le nombre des distances à recalculer.

Pour montrer la légitimité de cet algorithme il suffit de montrer que, dans l'algorithme usuel, les agglomérations successives, de niveau inférieur à la distance qui sépare deux voisins réciproques, ne modifient pas la propriété de ces deux points d'être l'un pour l'autre le plus proche voisin.

Soient k et k' une paire de voisins réciproques, et i et i' la paire à fusionner à l'étape considérée. Il n'est pas possible d'avoir à agréger i et k, par exemple, car d(i, k) > d(k, k'), donc d(i, k) n'est pas la plus petite des distances. En supposant que la formule de recalcul satisfait à la condition (6.3), on a après fusion :

$$d(iUi', k) > \mathrm{Min}\ (d(i, k)\ ,\ d(i', k))$$

mais comme k et k' sont voisins réciproques on a

$$d(k, k') < d(i, k)$$
$$d(k, k') < d(i', k)$$

il en résulte que

$$d(k, k') < d(iUi', k)$$

on montrerait de même que

$$d(k, k') < d(iUi', k')$$

Autrement dit la création du groupe iUi' ne change pas le fait que d(k, k') est la plus petite des longueurs des segments issus de k ou de k'. Ainsi, au fur et à mesure que se déroule l'algorithme élémentaire, les niveaux d'agrégation augmentent, jusqu'à ce que d(k, k') soit à son tour la plus petite des distances.

En résumé la hiérarchie du moment d'ordre deux peut se calculer en suivant l'algorithme suivant :

1) Pour chaque objet i rechercher son plus proche voisin que nous appellerons PPV(i)
2) Agréger toutes les paires de voisins réciproques c'est à dire les couples (i, i') tels que PPV(i) = i' et PPV(i') = i
3) Retourner en 1) tant que le nombre de groupes restants est supérieur ou égal à deux.

Il faut noter que les résultats sont rigoureusement identiques à ceux que l'on obtient par l'algorithme, maintenant traditionnel, des agrégations successives, tel qu'il a été décrit au chapitre 4.

3.- Application à l'exemple PSYSOC.

Les coordonnées des pays sur les trois premiers facteurs de l'AFC ont, encore une fois, servi de données pour l'algorithme des voisins réciproques ; celui-ci a été programmé (voir ci-dessous paragraphe 4) pour évaluer les distances selon la métrique euclidienne usuelle, tandis que les agrégations sont faites selon le critère du moment d'ordre deux. Les résultats sont très largement concordants avec les méthodes employées jusqu'ici (figure 1). On retrouve les trois groupes principaux déjà déterminés. Seules changent les subdivisions du grand groupe baptisé "Europe Nord".

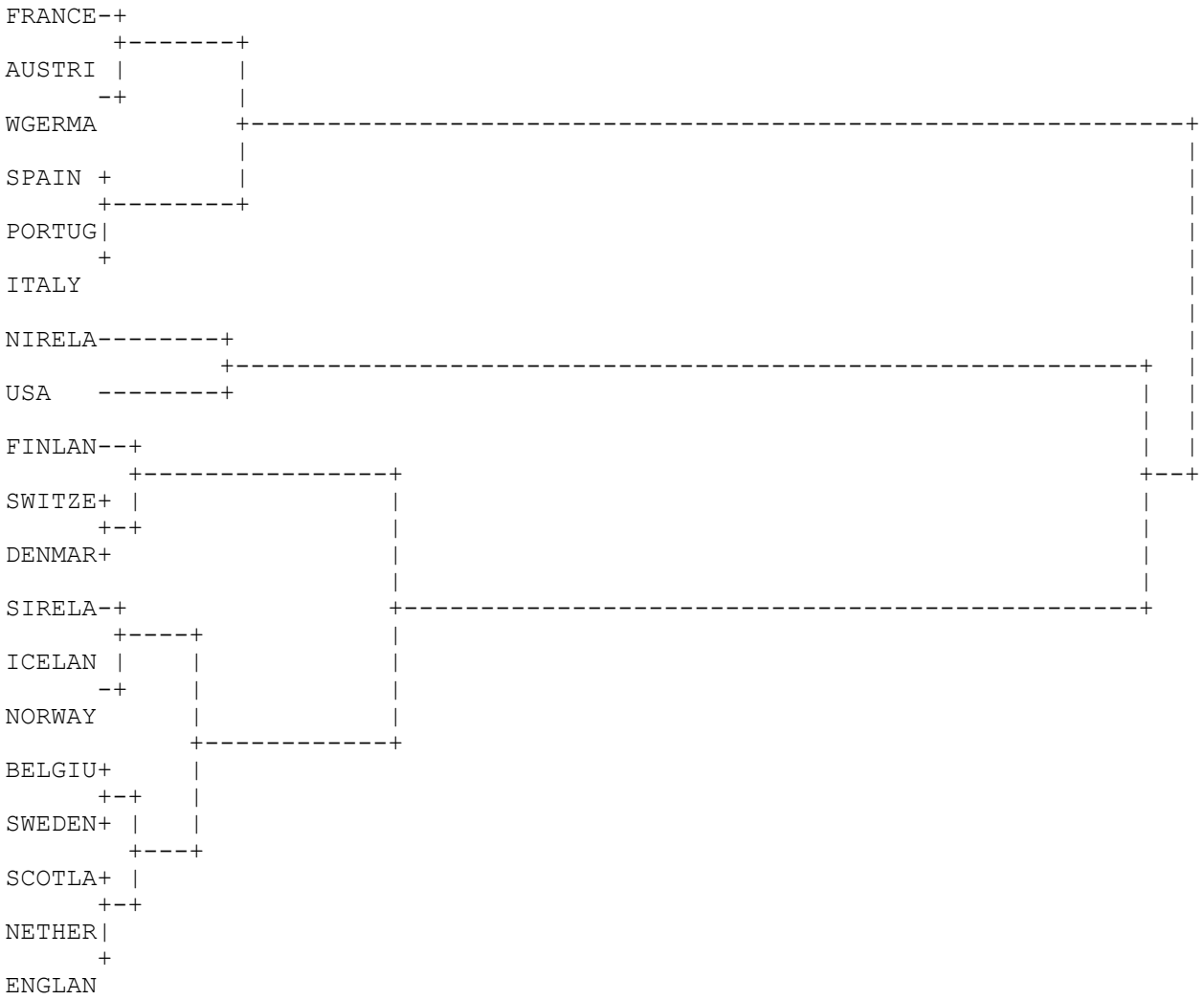


Figure 1 .- *Données PSYSOC, hiérarchie du Moment d'ordre deux calculée d'après les coordonnées factorielles (3 facteurs, A.F.C). Certains pays, par exemple AUSTRI et WGERMA, semblent ne pas être connectés à l'arbre ; ceci est du au fait que les niveaux de la hiérarchie sont très proches les uns des autres dans les faibles valeurs, et il n'est pas possible de les représenter sans distordre l'échelle globale de l'arbre.*

Il faut remarquer que, dans l'affichage des résultats, les niveaux d'agrégation des nœuds ne vont pas toujours en croissant. Cela résulte du principe même de l'algorithme dans lequel ceux-ci sont formés dès que l'on découvre des voisins réciproques, sans tenir compte de leur distance mutuelle. Les niveaux les plus hauts présentent entre eux de grands écarts par rapport aux niveaux inférieurs, ce qui semble indiquer que les groupes sont bien individualisés et homogènes. Cet aspect très tranché de l'arbre hiérarchique est trompeur. En effet les niveaux, ici, ne s'interprètent pas comme des distances mais comme des dispersions, ou, plus exactement, des accroissements de dispersion (voir ci-dessus, paragraphe 1). L'expérience montre aussi, et pour la même raison, que cette méthode tend à créer des groupes d'effectifs équilibrés.

4.- Procédure de calcul

Le classeur « AnaDon.xls » comporte la procédure *CAHmom2* qui réalise la construction hiérarchique du Moment d'ordre 2.

Chapitre 7

Construction descendante d'une hiérarchie

1.- Introduction

Les algorithmes de construction hiérarchique par agglomérations successives ou Constructions ascendantes hiérarchiques (CAH) sont, à juste titre, les plus couramment utilisés. Ils sont, en effet, rapides et l'expérience montre qu'ils fournissent des résultats cohérents. Cependant leur mode de fonctionnement par agrégations successives à partir des objets simples, suggère que les nœuds les plus élevés de la hiérarchie sont probablement peu représentatifs. Malheureusement c'est généralement sur eux que repose l'interprétation des résultats ; en effet l'utilisateur interprète habituellement la hiérarchie obtenue en examinant l'arbre réduit à ses seules branches principales.

Les méthodes basées sur des dichotomies successives, ou Constructions descendantes hiérarchiques (CDH), seraient plus satisfaisantes à cet égard. Ces méthodes partent de l'ensemble entier de tous les objets ; celui-ci est scindé en deux parties qui sont à leur tour scindées en deux, etc...jusqu'à ce que tous les sous ensembles obtenus soient réduits à un objet unique. Cependant ce type d'algorithmes a eu peu de succés jusqu'à présent à cause des inconvénients majeurs qu'il présente. En effet pour obtenir de bons résultats, il faudrait examiner à chaque étape toutes les dichotomies possibles pour n'en retenir qu'une, celle qui optimise un critère fixé à l'avance. Mais la scission en deux d'un groupe à n objets demande l'examen de 2^{n-1} - 1 bipartitions, ce qui requiert des calculs prohibitifs comme l'avaient déjà remarqué Edwards et Cavalli-Sforza dès 1965 (sans fournir de solution).

Même si l'examen d'un aussi grand nombre de bipartitions était techniquement réalisable, la hiérarchie obtenue n'optimiserait pas pour autant un critère global d'ajustement aux données ; mais les dichotomies ainsi obtenues pourraient sans doute être plus facilement interprétables. Pour éviter l'examen exhaustif de toutes ces dichotomies les auteurs de tels algorithmes ont eu recours à des simplifications que nous regroupons en trois grandes catégories :

- méthodes basées sur le choix ou la construction d'une variable particulière
- méthodes basées sur un ou plusieurs individus formant les embryons des sous ensembles
- méthodes basées sur la théorie des graphes

Bien que les méthodes utilisant la théorie des graphes soient très en vogue actuellement (Juin 2006) elles nécessitent quelques développements qui dépassent le cadre de cet ouvrage. Nous nous limiterons ici aux deux premières catégories de méthodes. Un autre problème ennuyeux réside dans le calcul des niveaux de jonction entre les branches de la hiérarchie ; selon la formule utilisée ces niveaux peuvent présenter des inversions, rendant impossible la représentation de l'arbre hiérarchique associé à la classification obtenue.

2.- Méthodes basées sur une variable particulière

Ces méthodes reposent sur le choix, ou sur la construction, d'une variable, que nous appellerons variable-critère. Cette variable, qui change à chaque étape, sert ensuite à effectuer la dichotomie. Supposons que l'on veuille scinder la classe C en deux sous-classes C' et C". Cette dichotomie se fera en mettant dans C' tous les objets présentant pour la variable-critère une valeur inférieure ou égale à un certain seuil et de ranger dans C" le reste des objets, c'est à dire ceux dont la valeur est supérieure au seuil choisi.

2.1.- Utilisation de l'une des variables des données

Le prototype de ce type d'algorithme est la méthode de Williams et Lambert (1959) que nous décrivons maintenant. Cette méthode est particulièrement rudimentaire. N'opérant que sur des variables qualitatives, elle sélectionne l'une des variables pour servir de critère d'affectation : tous les objets présentant, pour cette variable, la même modalité sont rangés dans la même classe (si les variables sont à plus de deux modalités le nœud correspondant aura plus de deux branches). La variable retenue est celle qui, dans la classe C à scinder, est la plus corrélée à toutes les autres. Comme il s'agit de variables qualitatives la corrélation est mesurée par le Khi-deux de contingence. On calcule donc les Khi-deux de contingence de toutes les variables prises deux à deux, et l'on retient celle pour laquelle la somme de ses Khi-deux est maximum.

La méthode de Williams et Lambert est bien adaptée au traitement de tableaux présentant un grand nombre d'observations et peu de variables qualitatives, ou questions. La table des Khi-deux de contingence entre variables est alors rapide à obtenir, par comparaison au temps qu'il faudrait pour calculer, par exemple, la matrice de Jaccard relative aux individus. En outre, à chaque nœud de la hiérarchie est attaché, par construction, le nom d'une variable, ce qui facilite l'interprétation : tous les individus associés à une même branche "répondent" de la même façon à toutes les questions (variables) qui ont abouti à la création de cette branche. Malheureusement les résultats sont en général grossiers. Cela tient au fait que les groupes d'individus se définissent rarement par leurs réponses strictement identiques à une série de questions mais bien plutot par un pourcentage élevé de réponses semblables sur l'ensemble des questions. Notons encore que les niveaux des nœuds de la hiérarchie ne sont plus définis que par l'ordre dans lequel ils apparaissent et il n'est pas naturel de leur associer un indice montrant la cohésion du groupe d'objets associés à ce nœud..

On pourrait imaginer un programme semblable travaillant sur des variables quantitatives. Il y faudrait ajouter une étape supplémentaire : une fois choisie la variable de scission, il faudrait choisir une valeur-seuil pour cette variable ; en dessous de ce seuil les objets seraient rangés dans l'une des sous-classes, au dessus de ce seuil les objets seraient affectés à l'autre sous-classe. Toutefois une telle procédure présenterait les mêmes avantages et les mêmes inconvénients que celle de Williams et Lambert.

2.2.- Utilisation des directions principales, ou axes factoriels

Plusieurs auteurs ont proposé des méthodes de ce type ; citons, par exemple, Reinert (1983), Boley (1998) et Chavent et al.(1999). Le principe général consiste à calculer, pour les seuls objets de la classe C à scinder, la première direction principale de ce sous-ensemble. Cette direction est la première composante principale s'il s'agit de variables quantitatives continues ou bien le premier axe factoriel de l'Analyse des Correspondances si les variables initiales sont qualitatives ou si elles représentent des comptages homogènes.

En général les coordonnées des objets sur les directions principales sont centrées de sorte que l'origine constitue le seuil naturel comme point de scission : les objets de coordonnées négatives sont mis dans l'une des sous-classes, ceux de coordonnées positives sont affectés à l'autre sous-classe. Il est possible, cependant, d'adopter une autre valeur-seuil pour optimiser, par exemple, la variance inter-classe.

Les résultats obtenus par de telles méthodes sont évidemment meilleurs que ceux de l'algorithme de Williams et Lambert, puisque les axes factoriels synthétisent généralement plusieurs variables. Elles sont efficaces pour le traitement du même format de tableau : nombreux individus mais peu de variables. En effet si les variables sont nombreuses alors le temps de calcul nécessaire à l'extraction du premier axe s'allonge rapidement. Par ailleurs, ces méthodes ne permettent pas d'associer une

variable à chaque noeud de la hiérarchie, puisque ceux-ci sont définis par une combinaison linéaire des variables initiales.

3.- Méthodes basées sur des individus particuliers

3.1. Sélection d'un point périphérique : méthode de McNaughton-Smith et al.(1964).

Pour initier une dichotomie ces auteurs examinent la somme des distances de chaque objets à tous les objets de sa propre classe. Celui dont la somme des distances est maximum est supprimé de sa classe et est pris comme embryon, ou noyau, d'une nouvelle classe. Appelons encore C la classe qui perd cet élément et C' la nouvelle classe. La suite de l'algorithme consiste à transférer un à un certains éléments de C vers C' de façon à optimiser un critère local. L'article de McNaughton-Smith et al. ne donne aucune précision sur ce critère mais on peut penser à maximiser la variance interclasse, pour les deux classes C et C', ou bien la distance moyenne inter-classe. La procédure de transfert est arrètée lorsque le critère cesse de s'améliorer.

3.2. Sélection de deux points périphériques : méthode de Hubert (1973).

La méthode de Hubert diffère de la précédente en ce que les scissions successives sonr initiées par les deux points les plus éloignés de la classe à scinder. Hubert a proposé diverses variantes de sa méthode qui ne diffèrent entre elles que par le mode d'affectation qui est toujours basé sur les distances aux deux points les plus éloignés de la classe à scinder. Ainsi dans la variante la plus élémentaire, si un objet est plus proche du premier que du second de ces points il est mis dans la première classe. Sinon il est affecté à l'autre classe. Les autres variantes consistent à examiner les distances rangées par ordre croissant. On n'affecte alors un objet à une sous-classe que s'il est suffisammnent proche de l'un, ou de tous les objets déjà affectés à cette sous-classe.

Les résultats obtenus par l'un ou l'autre des algorithmes de Hubert ne donnent pas non plus satisfaction. Nous pensons que l'affectation basée sur les distances aux points les plus éloignés est discutable. En effet, ces points sont souvent des observations accidentelles, voire aberrantes, et en tous cas non représentatives des grandes masses de la classe considérée. De sorte que la dichotomie qui en résulte ne représente pas correctement la répartition des objets de la classe.

3.3. Sélection de deux points noyaux : méthode de Roux (1995)

Soit q un sous-ensemble de l'ensemble I des objets à classer. On examine un certain nombre de partitions de q en 2 classes, ou bipartitions ; on dit qu'une bipartition est induite par i et i' (tous deux éléments de q) si elle est formée de la façon suivante : C(i) est le sous-ensemble de q dont tous les éléments sont plus proches de i que de i', et de façon analogue, C(i') a tous ses éléments plus proches de i' que de i.

Le critère pour décider qu'une classe q sera scindée en deux est basé sur la distance moyenne inter-classe :

$$M(q,i,i') = (1/(n_i n_{i'})) \Sigma_{k \in C(i),\, k' \in C(i')} d_{kk'}$$

Dans cette formule n_i et $n_{i'}$ désignent les effectifs des deux groupes C(i) et C(i') respectivement. L'algorithme se déroule comme suit.

a) Mise à l'état initial.
Au début tous les objets appartiennent à la même classe.

b) A chaque étape on a une partition Q de l'ensemble des objets. Pour toutes les classes q de Q, d'effectif supérieur ou égal à 2, on calcule le critère :

$$Crit(q) = Max_{\,i,i' \in q}\, M(q,i,i')$$

c) On subdivise la classe q^* qui maximise ce critère :

$$Crit(q^*) = Max_{\,q \in Q}\, Crit(q)$$

d) S'il reste des classes à 2 éléments ou plus on retourne en b) sinon on arrête.

Dans une version précédente de ce travail (Roux, 1985) nous avions envisagé un critère de scission basé sur la variance des distances inter-groupe (comme Edwards and Cavalli-Sforza, 1965, ou Fages, 1978), mais les résultats obtenus étaient de qualité moyenne et nous avons abandonné ce critère.

A chaque scission cette façon de procéder conduit à examiner, au plus, n(n-1)/2 partitions, au lieu des 2^{n-1} - 1 bipartitions possibles. Le calcul du critère est lui-même d'ordre n ; enfin le nombre total de scissions à effectuer est égal à n-1. On a donc un algorithme de complexité polynomiale de degré 4. C'est un ordre élevé mais qui reste réalisable avec les ordinateurs actuels. En accord avec le critère de scission, il est naturel de fixer les niveaux de la hiérarchie égaux à la distance moyenne entre les groupes qu'ils définissent. C'est pourquoi nous avons appelé CDH-LM le programme réalisant cet algorithme.

4.- Le problème des inversions

La procédure ci-dessus présente un grave inconvénient : elle ne garantit pas contre l'apparition d'inversions dans la hiérarchie, laquelle est alors impossible à construire, et jette quelques doutes sur sa validité. Ce phénomène bien que peu fréquent (environ 10 % des cas selon nos essais), demande un aménagement de la méthode. Pour cela plusieurs stratégies sont possibles.

La première stratégie consisterait à simplement signaler, par un message à l'utilisateur, qu'une inversion s'est produite. La seconde stratégie possible est celle adoptée par Kaufman et Rousseeuw (1990) : les niveaux d'agrégations sont les diamètres des classes correspondantes. Comme les sous-classes sont nécessairement d'un diamètre inférieur ou égal à la classe qui les englobe, il ne peut y avoir d'inversion.

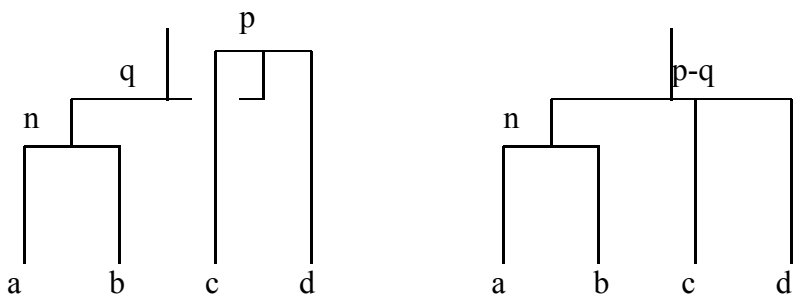


Fig. 1, En cas d'inversion les deux noeuds concernés, p et q, sont fusionnés et leur niveau est calculé selon la distance moyenne entre les trois groupes n, c et d.

Nous proposons une troisième stratégie dans laquelle, après la construction descendante, on contrôle les niveaux des partitions successives. Dans le cas où l'on découvre une inversion on recalcule la distance moyenne entre les trois groupes concernés, et cette distance moyenne est prise comme niveau commun aux deux noeuds en inversion.

5.- Applications aux exemples

5.1. Exemple PSYSOC

Nous avons traité le tableau des distances relatif aux données PSYSOC par l'algorithme que l'on vient de décrire (CDH-LM, paragraphe 3.3). Les distances sont calculées par la formule euclidienne usuelle sur les 3 premiers axes de l'A.F.C. Les résultats (voir figure 1) sont comparables à ceux que procure la hiérarchie ascendante de la distance moyenne, quoique les deux pays excentriques, Irlande du Nord (NIRELA) et USA, tout en étant isolés des autres, ne soient pas mis dans un même groupe. Mais on y retrouve bien le groupe "Méditerranéen" (SPAIN, ITALY et PORTUG) relié au groupe « Europe moyenne » (FRANCE, WGERMA et AUSTRIA). Les autres pays sont ceux du groupe "Europe-Nord" dans lequel il est difficile de discerner des sous-groupes.

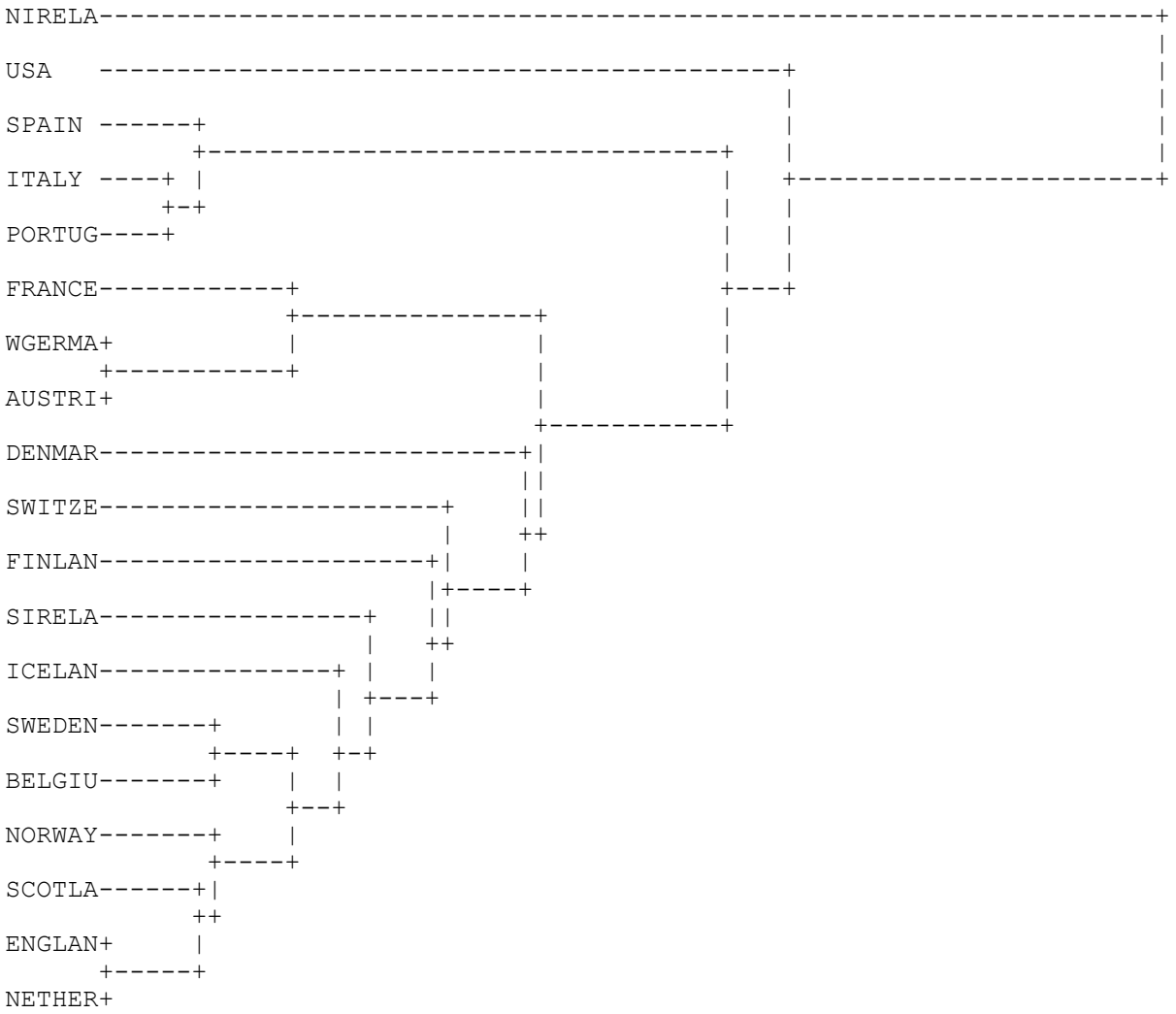


Fig. 2. *Données PSYSOC. Hérarchie obtenue par l'algorithme CDH-LM de construction descendante selon le lien moyen. Les données de base sont les distances entre les pays, calculées d'après les coordonnées factorielles issues de l'A.F.C. (3 premiers facteurs).*

5.2. Exemple PHYTOS

L'algorithme CDH-LM a été appliqué aux données phytosociologiques, en partant de la distance de Jaccard entre relevés. Les résultats (voir figure 2) ne concordent pas avec ceux que fournissent les algorithmes élémentaires de construction ascendante (voir chapitre 4, paragraphe 2.2). Seul le groupement (3, 4, 14, 16) (Pacn1 : Festucetum halleri, subass. nardetosum, faciès normal) apparaît clairement. Les autres relevés sont mélangés et l'on n'y reconnaît aucun des groupements identifiés.

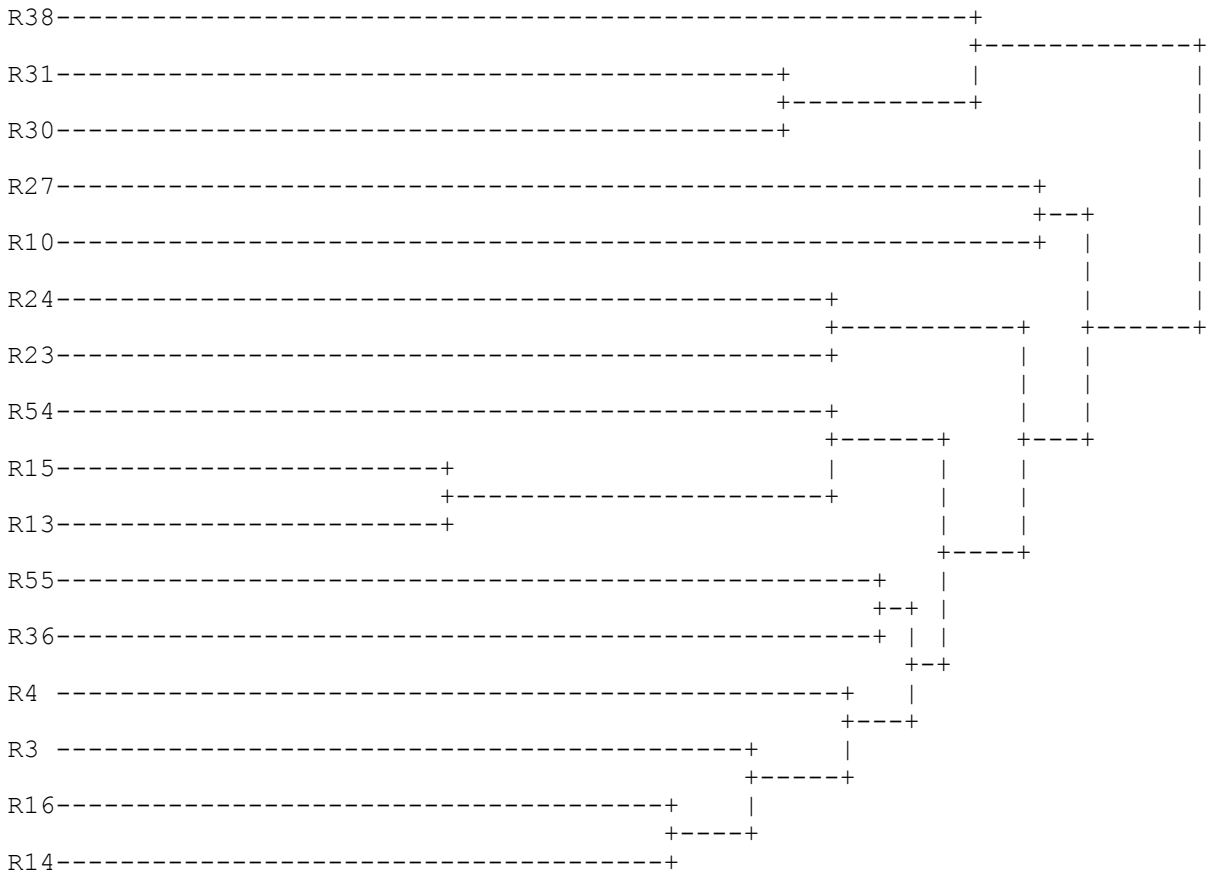


Fig. 3. Données PHYTOS. Hiérarchie obtenue par l'algorithme CDH-LM de construction descendante, à partir de la matrice des distances de Jaccard

6.- Conclusion

Les constructions hiérarchiques par divisions successives ont un aspect séduisant : elles commencent par le haut de l'arbre, c'est à dire par la partie sur laquelle repose essentiellement l'interprétation. Malheureusement les simplifications drastiques qu'elles exigent, pour maintenir des temps de calcul raisonnables, font que les résultats obtenus sont souvent décevants. Cependant les dichotomies basées sur des variables bien choisies ont l'avantage d'être rapides et de fournir des interprétations aisées. Elles permettent donc de traiter facilement de très grands jeux de données avec peu de variables.

7.- Procédure de calcul.

La procédure *CDHLM*, dans le classeur « AnaDon.xls » réalise la construction descendante décrite au paragraphe 3.3.

Chapitre 8

Aides pour l'interprétation des classifications

Lorsque, par l'une des méthodes des chapitres précédents, on a obtenu une classification des objets, on souhaite, en général, savoir quelles sont les variables responsables de tel ou tel regroupement. C'est ce problème que l'on va étudier dans le présent chapitre en séparant, comme il se doit, le cas de variables quantitatives de celui des variables qualitatives.

1.- Variables quantitatives.

On a vu au chapitre 5, équation (5.3), que le moment d'ordre deux d'un solide peut se décomposer en moment intra-classe et moment inter-classe. Ce dernier, qui est ce qu'on appelle le moment d'ordre deux d'une partition, représente la dispersion des centres de gravité, dans laquelle on tient compte des masses, c'est à dire des effectifs des classes. Le rôle des variables peut être facilement apprécié dans leur contribution à cette dispersion. Comme au chapitre 5 on suppose que la distance utilisée est la distance euclidienne usuelle.

1.1.- Interprétation d'une partition

Reprenant les notations du chapitre 5, on appelle Q la partition formée des classes q, q' ..., d'effectifs m_q, $m_{q'}$ ..., dont les centres de gravité sont g_q, $g_{q'}$, ... Le moment d'ordre deux de la partition Q est :

$$M^2(Q) = \Sigma_q \, m_q \, d^2(g_q, g)$$

où g (sans indice) désigne le centre de gravité de l'ensemble de tous les objets. Le carré de la distance euclidienne entre g_q et g s'écrit :

$$d^2(g_q, g) = \Sigma_{j \in J} \, (g_q(j) - g(j))^2$$

où J représente l'ensemble des variables, g(j) est la j-ème coordonnée du point g. En intervertissant l'ordre de sommation on a donc :

$$M^2(Q) = \Sigma_{j \in J} \, \Sigma_q \, m_q \, (g_q(j) - g(j))^2 \qquad (8.1)$$

On appellera contribution de la variable j à la classe q, la quantité :

$$CTR(q, j) = m_q \, (g_q(j) - g(j))^2$$

Remarquons que cette quantité est toujours positive ; cependant il peut être utile de connaitre le signe de la différence entre parenthèses, pour savoir si la variable j est inférieure ou supérieure à la moyenne générale, g(j) , dans la classe considèrée.

Dans la présentation des résultats on publiera deux tableaux. Le premier s'appelle "contributions des variables aux classes" et donne les quantités ci-dessus, munies du signe convenable, exprimées en pourcentage, relativement à la dispersion de la classe, c'est à dire à la somme de ces quantités pour toutes les variables, la classe étant fixée.

Le second tableau, dénommé "contributions des classes aux variables" fournit, en pourcentages également, le rapport de la contribution à la dispersion de chaque variable, c'est à dire à la somme des contributions pour toutes les classes et pour une variable fixée.

Si l'on s'intéresse à l'interprétation des classes c'est donc le premier tableau qu'il faudra examiner. Au contraire si une ou plusieurs variables ont un rôle important il vaudra mieux étudier le second tableau.

1.2.- Interprétation d'une hiérarchie

Lorsqu'on a établi une hiérarchie sur un ensemble I d'objets on désire, en général, savoir quelles sont les variables de l'ensemble J, déterminantes pour la formation de chaque nœud de l'arbre. Dans le cas de variables quantitatives, comme précédemment, on examine le rôle joué par chaque variable dans le carré de la distance $d^2(g_q, g_{q'})$ entre les centres de gravité des deux classes q et q' constitutives de chaque nœud :

$$d^2(g_q, g_{q'}) = \Sigma_{j \in J}(g_q(j) - g_{q'}(j))^2 \qquad (8.2)$$

C'est donc la quantité $(g_q(j) - g_{q'}(j))^2$ qu'on appelle contribution de la variable j au nœud considéré. Et le programme de calcul fournira un tableau dont les lignes sont les nœuds successifs de la hiérarchie et dont les colonnes représentent les variables. Dans ce tableau les contributions seront rapportées à leur somme pour toutes les variables et exprimées en pourcentage de cette somme.

2.- Variables qualitatives

Dans le cas de variables qualitatives le calcul du centre de gravité n'aurait pas de sens. On ne peut donc pas utiliser les formules du paragraphe précédent. En revanche on dispose d'un critère bien adapté à notre problème : la formule du Khi-deux de contingence entre deux variables. On peut, en effet, considérer une partition en k classes comme une variable qualitative à k modalités ou états. Le Khi-deux de contingence entre une variable et les classes d'une partition indique le degré de liaison de cette variable avec la partition.

Dans le cas d'une hiérarchie on considérera le rôle des variables nœud par nœud, un nœud étant considéré comme une variable qualitative à deux catégories ; en effet tout objet de la classe associée au nœud appartient à l'une ou à l'autre des sous-classes associées aux deux branches réunies. Rappelons la formule du Khi-deux ; cette quantité est égale à la somme des carrés des écarts entre effectifs observés et effectifs théoriques, pondérés par les effectifs théoriques :

$$\text{Khi-2} = \Sigma \ (\text{effectifs observés} - \text{effectifs théoriques})^2 \ / \ \text{effectifs théoriques}$$

Dans le cas d'un tableau de contingence, où les effectifs e_{kl} se répartissent dans un tableau, dont les lignes sont indicées par la lettre k et les colonnes par la lettre l, cette formule devient :

$$\text{Khi-2} = \Sigma_k \ \Sigma_l (e_{kl} - e_{k.}\ e_{.l}\ /m)^2 \ / \ (e_{k.}\ e_{.l}\ /m) \qquad (8.4)$$

où m est l'effectif total des objets. On appelle $e_{k.}$ l'effectif de la modalité k, tandis que e_{l} est l'effectif de la classe l.

2.1.- Interprétation d'une partition

Dans le cas d'une partition on demande à l'ordinateur de dresser un tableau [variables * classes-partition], où l'on trouve à l'intersection de la ligne j et de la colonne k la valeur CTR(j, k) de la contribution de la variable j à la classe k de la partition. Pour cela il suffit d'effectuer, dans la double somme ci-dessus (8.4), la partie relative aux classes l de la variable considérée :

$$\text{CTR}(j,k) = \Sigma_{l \in L(j)} \ (e_{kl} - e_{k.}\ e_{.l}\ /m)^2 \ / \ (e_{k.}\ e_{.l},\ /m)$$

où l'indice l parcourt l'ensemble L(j) des modalités de la variable j. Il est clair que la somme de ces nombres, obtenue en faisant varier k sur l'ensemble des classes de la partition, est égale au Khi-2. Pour plus de commodité ces nombres sont divisés par leur somme et sont exprimés en millièmes, de façon à déterminer facilement les classes les mieux caractérisées par la variable j étudiée. Il faut noter que, dans le cas d'une variable j à deux modalités, comme la présence ou l'absence d'une espèce, une classe peut être caractérisée aussi bien par la présence que par l'absence de l'espèce en question.

Le tableau est complété par la valeur du Khi-deux, et par le nombre de degrés de liberté à prendre en compte dans une éventuelle procédure de test statistique.

2.2.- Interprétation d'une hiérarchie

Pour aider au dépouillement d'une hiérarchie, on dresse un tableau [variables * nœuds], donnant les contributions CTR(j, n) de la variable j à l'écart entre les deux classes formant le nœud n. La formule (8.4) fournit encore les valeurs cherchées mais l'indice k n'y peut prendre que deux valeurs correspondant aux deux classes en question. L'indice l représente, comme précédemment, les classes de la variable considérée. Et les effectifs ne prennent en compte que les objets appartenant au nœud n :

$$CTR(j,\ n) = \Sigma_{k \in \{an,\ bn\}} \quad \Sigma_{l \in L(j)} \ (e_{kl} - e_{k.}\ e_{.1}\ /m)^2 \ /\ (e_{k.}\ e_{.1},\ /m)$$

Dans cette formule {an, bn} désigne l'ensemble à deux éléments, formé de l'aîné et du benjamin du nœud n. Il faut aussi prendre garde que l'effectif m est ici le nombre d'objets impliqués dans le nœud n, et non l'effectif total de tous les objets étudiés.

3.- Application aux exemples

3.1 .- Données PSYSOC (quantitatives)

On a effectué les calculs de contributions en utilisant d'abord la partition en trois classes que nous connaissons bien :

```
Classe 1 : AUSTRI , FRANCE , WGERMA, ITALY, SPAIN, PORTUG
Classe 2 : BELGIU, SWEDEN, SCOTLA, NETHER, ENGLAN, ICELAN,  NORWAY, SIRELA,
FINLAN, SWITZE, DENMAR
Classe 3 : NIRELA, USA
```

Il faut noter, en passant, que la façon dont cette partition a été obtenue importe peu ; on recherche simplement quelles sont les variables initiales les plus caractéristiques de chaque classe. C'est pourquoi le moment d'ordre deux total et le moment inter-classe figurant au tableau ci-dessous, calculés sur ces variables, ne coïncident pas avec les quantités homologues

que l'on a pu obtenir avec l'algorithme des centres mobiles appliqué aux coordonnées factorielles des pays.

```
MOMENT TOTAL = 556228
MOMENT INTERCLASSE = 261834         R = 0.47

CONTRIBUTIONS VAR. /CLASSES

   SUI   HOM   ARO   AIN   AAU   CIR

1    O     O     8     O     0    91
2    1    -2   -12     O     O   -85
3  -18    63     2     O    -2   -15
```

Tableau 1. *Données PSYSOC, contributions des variables aux classes*

```
CONTRIBUTIONS CLASSES/VAR.

   SUI   HOM   ARO   AIN   AAU   CIR

1    O    -3    58     O    38    68
2   17    -8   -39   -13    -2   -30
3  -83    89     2    87   -60     2
```

Tableau 2. *Données PSYSOC, contributions des classes aux variables*

Il ressort nettement du tableau 1 que la classe 3 se caractérise principalement par un taux élevé d'Homicides alors que les Suicides et les Cirrhoses du foie y sont à un niveau inférieur à la moyenne (signes négatifs). Ce qui caractérise de façon quasi exclusive les classes 1 et 2 ce sont les Cirrhoses du foie, en quantité importante dans la classe 1, excessivement peu nombreuses (signe négatif) dans la classe 2. Dans une moindre mesure ces deux classes se différencient également par les Accidents de la route, nombreux dans la classe 1, plus rares dans la classe 2.

Le tableau 2 fournit des renseignements intéressants sur la dispersion de chaque variable relativement aux classes. Ainsi on peut dire que les Cirrhoses sont à des taux très voisins les uns des autres pour les pays de la classe 1, alors que ces taux sont plus dispersés pour la classe 2. Autrement dit les taux élevés de Cirrhose sont beaucoup plus caractéristiques de la classe 1 que ne le sont les taux faibles pour la classe 2.

Nous allons examiner maintenant la hiérarchie ascendante de la distance moyenne, obtenue on prenant pour distance initiale la distance du Khi-deux sur les données brutes. Les contributions des variables aux nœuds de la hiérarchie sont décrites au tableau 3, dont les lignes représentent les nœuds de la hiérarchie, et dont les colonnes sont les variables.

	SUI	HOM	ARO	AIN	AAU	CIR
20	5	0	53	3	22	17
21	5	0	76	1	11	6
22	40	7	24	1	1	27
23	7	0	19	6	56	12
24	2	1	26	0	31	41
25	10	0	23	0	47	20
26	1	0	43	0	1	55
27	36	2	9	3	0	49
28	4	0	75	0	7	14
29	57	0	10	2	30	0
30	1	1	1	0	93	4
31	7	0	5	1	70	16
32	9	0	6	1	82	2
33	46	0	2	0	52	0
34	88	0	0	0	0	12
35	28	5	0	0	10	57
36	0	0	9	0	0	90
37	18	63	2	0	2	15

Tableau 3.- *Contribution des variables aux nœuds de la Hiérarchie ascendante du lien moyen.*

La difficulté pour interpréter ce tableau provient de ce qu'il est nécessaire d'identifier les nœuds. Pour cela il faut se reporter à la description de la hiérarchie telle qu'elle figure au chapitre 4. En fait, seuls les nœuds les plus hauts de la hiérarchie sont réellement utiles, c'est pourquoi nous n'avons reconstruit que la partie supérieure de l'arbre avec les numéros des nœuds (figure 1).

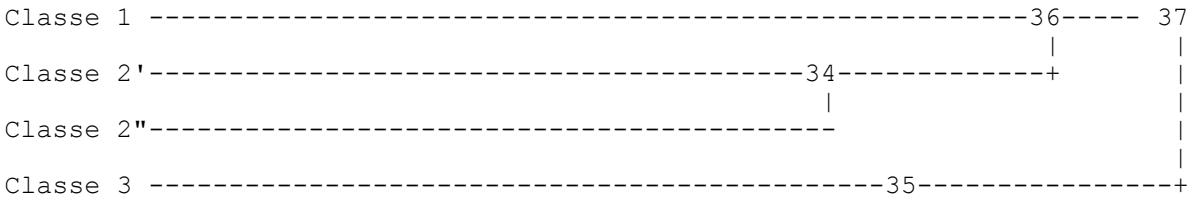


Figure 1.- *Partie supérieure de l'arbre hiérarchique de la distance moyenne (Données PSYSOC, distance du Khi-2 sur données brutes).*

Dans cette figure on a dissocié la classe 2 en ses deux sous classes :

```
Classe 2' : FINLAN, SWEDEN, SWITZE, DENMAR
Classe 2" : BELGIU, NETHER, ENGLAN,ICELAN, SCOTLA, NORWAY, SIRELA
```

Examinons d'abord le dernier nœud (37) qui relie la classe 3 (NIRELA et USA) aux deux autres. La dernière ligne du tableau 3 montre clairement que ce sont les Homicides qui départagent ces deux groupes de pays. Le nœud 36 relie la classe 1 et la classe 2. Il est caractérisé par la variable Cirrhose du foie qui explique 90 % de la dispersion interclasse. Ces renseignements ne font que confirmer ceux que nous avions déjà recueillis par l'observation des contributions aux classes de la partition. Mais on peut continuer avec l'examen des nœuds suivants. En particulier le nœud 34 attire notre attention sur les deux sous-classes 2' et 2" décrites ci-dessus. On voit que ce sont les Suicides qui, cette fois-ci, permnettent de distinguer ces deux sous-classes. Un coup d'oeil au tableau des données montre qu'en effet, les pays de la sous-classe 2' ont des taux de suicides nettement plus élevés que la moyenne qui est de 132 ; les pays de l'autre sous-classe (2") ayant naturellement des taux inférieurs à la moyenne (sauf BELGIU).

3.2.- Données PHYTOS (qualitatives).

Pour cette application nous reprenons la partition en 4 classes mise en avant au chapitre 2 (paragraphe 1.2) et retrouvée dans les applications précédentes ; le relevé 24, d'affectation incertaine, a été attribué au groupement Pacn1, comme cela a été proposé au chapitre 2.

CL.1	Groupement Pan	relevés 13,15.23	Nardetum alpigenum
CL.2	Groupement Pacnl	relevés 3,4,14,16,24	Festucetum halleri subass. Nardetosum (faciès normal)
CL.3	Groupement Pacn2	relevés 10,54,55	Idem mais faciès à *Elyna* et *Salix*
CL.4	Groupement Pac	relevés 27,30,31,36,38	Festucetun halleri sensu stricto

Le tableau 4 donne, pour toutes les espèces (en lignes), leur contribution aux quatre classes de cette "partition vedette". L'avant-dernière colonne représente la somme de ces contributions, c'est à dire le degré de liaison globale entre l'espèce et la partition, égale au Khi-deux de contingence. Les contributions sont exprimées on millièmes. Comme les espèces sont toujours des variables à deux modalités (présence ou absence), tous ces nombres sont comparables d'une ligne à l'autre. Cependant on doit se souvenir que l'absence d'une espèce joue autant que la présence dans la valeur du Khi-deux, ce qui conduit à caractériser les classes plus souvent par l'absence que par la présence.

Il convient d'examiner d'abord la colonne Khi-deux pour déterminer les espèces les plus importantes. On remarque les espèces suivantes (valeur du Khi-deux entre parenthèses) :

```
117  Potentilla aurea (16)
131  Salix herbacea (16)
144  Sempervivum arachnoideum (12.5867)
129  Sagina glabra (Wild) (12.4444)
 72  Geum montanum (11.7333)
 82  Homogyne alpina (11.7333)
159  Trifolium pratense L. (11.7333)
```

Un retour sur le tableau des données (chapitre 2) nous permet de constater que l'espèce 117 est absente de la classe 4 et présente partout ailleurs. Au contraire, l'espèce 131 n'est présente que dans la classe 2. Ceci confirme que le critère, basé sur le Khi-deux attribue autant d'importance à l'absence qu'à la présence d'une espèce. L'espèce 144 est une caractéristique de la classe 4, quoiqu'elle apparaisse une fois ailleurs. De même pour l'espèce 129, caractéristique du groupe 3, mais qui apparaît une fois dans un autre relevé (le numéro 54). L'espèce 72 se distingue parce qu'elle est présente partout sauf dans la classe 4, encore que l'un des relevés de cette classe la possède. On pourrait continuer ainsi l'examen des espèces par ordre des coefficients de liaison décroissants. On voit que ces calculs auxiliaires font clairement apparaître les variables discriminant les différents groupes.

	CL.1	CL.2	CL.3	CL.4	TOTAL	KHI-2	D.D.L.
1	38	63	563	338	1000	1.7778	3
2	64	16	458	462	1000	8.4148	3
5	159	58	624	159	1000	7.2479	3
7	82	380	488	50	1000	3.5879	3
10	688	85	85	142	1000	5.0286	3
11	218	131	634	17	1000	10.4145	3
12	311	240	187	263	1000	7.3115	3
20	188	313	313	188	1000	1.3714	3
21	142	85	85	688	1000	5.0286	3
24	2	241	241	516	1000	7.4667	3
26	276	165	460	99	1000	6.0444	3
29	169	94	344	393	1000	6.7894	3
41	467	261	54	219	1000	2.4550	3
42	607	12	12	367	1000	1.7778	3
45	17	710	256	17	1000	3.9111	3
48	688	85	85	142	1000	5.0286	3
50	263	187	240	311	1000	7.3115	3
53	276	460	165	99	1000	6.0444	3
55	218	634	131	17	1000	10.4145	3
57	516	241	241	2	1000	2.8718	3
60	38	63	563	338	1000	1.7778	3
61	32	1	719	248	1000	9.1733	3
62	188	313	313	188	1000	0.0711	3
63	49	289	289	374	1000	6.0703	3
64	219	261	54	467	1000	2.4550	3
65	517	26	452	6	1000	5.1640	3
67	337	562	62	37	1000	1.7778	3
68	540	165	18	276	1000	6.0444	3
69	187	313	313	187	1000	9.6000	3
70	99	165	460	276	1000	2.5905	3
71	142	85	85	687	1000	8.1231	3
72	142	85	85	688	1000	11.7333	3
75	276	460	165	99	1000	6.0444	3
77	467	261	54	219	1000	2.4550	3
79	364	18	87	530	1000	7.3312	3
82	6	85	767	142	1000	11.7333	3
84	142	85	85	687	1000	8.1231	3
86	688	85	85	142	1000	5.0286	3
87	99	165	460	276	1000	2.5905	3
90	72	812	43	72	1000	9.9048	3
95	248	1	719	32	1000	9.1733	3
98	142	85	85	687	1000	8.1231	3
100	142	85	85	687	1000	8.1231	3
105	99	165	460	276	1000	2.5905	3
109	6	26	452	517	1000	5.1640	3
112	99	165	460	276	1000	2.5905	3
113	276	18	165	540	1000	6.0444	3
114	72	813	43	72	1000	4.6222	3
116	17	634	131	218	1000	10.4145	3
117	142	85	85	688	1000	16.0000	3
120	3	210	210	578	1000	11.1238	3
125	142	85	85	687	1000	2.3467	3
126	99	165	460	276	1000	6.0444	3
129	134	9	723	134	1000	12.4444	3
130	99	460	165	276	1000	2.5905	3
131	72	813	43	72	1000	16.0000	3
144	52	143	143	662	1000	12.5867	3
145	188	313	313	188	1000	3.2000	3
156	142	85	85	688	1000	5.0286	3
157	142	85	85	688	1000	5.0286	3
158	32	1	719	248	1000	9.1733	3
159	6	85	767	142	1000	11.7333	3
160	5	617	373	5	1000	1.1214	3
163	373	3	622	2	1000	8.0356	3
166	613	188	188	13	1000	5.3333	3
168	142	85	85	688	1000	5.0286	3

Tableau 4.- *Données PHYTOS, contribution des variables (espèces) aux classes d'une partition*

	N17	N18	N19	N20	N21	N22	N23	N24	N25	N26	N27	N28	N29	N30	N31
1	0	100	0	0	100	0	0	11	25	6	2	0	8	100	5
2	0	0	100	0	25	0	0	33	100	17	7	0	21	0	35
5	100	100	0	25	0	0	0	0	25	0	24	0	29	0	10
7	0	0	0	100	100	0	100	11	0	6	9	100	29	3	2
10	0	0	100	0	25	0	0	11	0	38	20	0	8	0	6
11	0	0	0	0	0	100	100	0	0	0	0	25	100	3	2
12	0	100	0	0	0	100	100	0	25	0	56	25	8	3	17
20	0	0	0	0	0	0	100	100	0	6	9	0	4	38	2
21	0	0	0	0	0	0	0	0	0	0	0	0	0	100	31
24	0	0	0	0	100	100	0	33	0	17	20	25	8	17	35
26	100	0	0	25	0	0	0	0	0	0	0	100	59	44	5
29	0	100	0	0	0	0	100	0	25	100	22	0	14	38	24
41	0	100	0	100	0	100	100	0	25	100	22	25	0	3	5
42	0	0	100	100	25	0	100	33	100	17	7	0	2	38	6
45	0	0	0	0	0	0	100	100	100	6	22	0	14	38	1
48	0	0	100	0	25	0	0	33	0	17	20	0	8	0	6
50	100	0	100	25	25	0	0	33	0	17	36	0	7	0	21
53	0	100	0	0	0	0	100	0	25	0	56	100	8	3	5
55	0	0	0	0	0	0	100	0	0	0	100	100	14	3	2
57	0	0	0	0	100	0	0	33	0	17	20	100	8	17	0
60	0	100	0	0	0	100	0	100	25	6	2	25	8	44	5
61	0	100	0	0	0	0	0	100	25	6	2	0	45	0	21
62	0	100	100	100	25	100	0	11	25	38	0	25	0	44	0
63	100	0	0	25	0	0	0	100	100	6	22	0	7	0	21
64	100	0	0	25	0	0	100	100	100	38	0	0	7	38	10
65	0	0	0	0	0	100	100	0	100	100	22	25	31	3	0
67	0	0	100	100	25	0	0	11	0	38	20	100	1	17	1
68	100	0	100	25	25	0	0	33	0	17	36	0	0	0	15
69	0	0	0	0	0	0	100	0	0	100	9	0	66	38	16
70	0	0	0	100	0	0	0	0	0	100	9	0	6	0	6
71	0	0	0	0	0	0	0	0	0	0	0	100	0	44	51
72	0	0	0	0	0	0	100	0	0	0	0	0	0	38	73
75	0	100	0	0	0	0	0	0	25	0	56	0	8	100	5
77	0	0	0	100	0	100	0	0	100	100	22	25	0	44	5
79	100	0	0	25	100	0	0	11	100	6	22	0	0	0	35
82	0	0	100	0	25	0	0	11	0	6	9	0	66	0	15
84	0	0	0	0	0	0	0	0	0	0	0	0	0	100	51
86	0	0	0	0	100	0	0	33	0	17	20	0	8	0	6
87	0	0	0	100	0	0	0	0	0	100	9	0	6	0	6
90	0	0	0	0	0	0	0	0	100	0	56	0	8	0	6
95	0	100	0	0	0	0	100	0	25	0	24	0	66	38	3
98	0	0	0	0	0	100	100	0	0	0	0	25	0	3	51
100	0	0	0	0	0	100	100	0	0	0	0	25	0	3	51
105	0	0	0	100	0	0	0	0	0	100	9	0	6	0	6
109	0	0	100	0	25	0	100	33	100	17	0	0	14	38	24
112	0	0	0	100	100	0	0	11	0	6	9	0	6	0	6
113	0	0	0	0	0	0	0	0	100	0	24	0	4	100	30
114	0	0	0	0	0	0	0	0	100	0	24	0	4	0	3
116	0	0	0	0	100	0	0	33	0	17	36	0	31	0	21
117	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
120	0	0	100	0	25	0	0	33	0	17	20	0	8	0	58
125	0	0	0	0	0	0	0	0	0	0	0	100	0	17	15
126	0	0	100	100	25	0	0	33	0	17	20	0	15	0	15
129	0	100	0	0	0	0	0	0	25	0	24	0	66	0	15
130	0	0	0	0	100	0	0	11	100	6	2	0	8	0	6
131	0	0	0	0	0	0	0	0	0	0	100	0	14	0	10
144	0	0	100	0	25	0	0	11	0	6	9	0	4	0	76
145	0	0	100	0	25	0	0	11	0	38	20	100	8	44	5
156	0	0	0	0	0	0	0	0	0	0	0	0	0	100	31
157	0	0	0	0	0	100	0	0	0	0	0	25	0	44	31
158	0	100	0	0	0	0	0	100	25	6	2	0	45	0	21
159	0	0	0	0	100	0	0	11	0	6	9	0	66	0	15
160	100	0	0	25	0	0	100	100	0	6	9	0	6	38	0
163	0	100	0	0	0	100	0	0	25	0	24	25	66	44	0
166	0	0	0	0	100	100	0	33	0	17	36	25	14	17	1
168	0	0	0	0	0	0	0	0	0	0	0	100	0	44	31

Tableau 5.- *Données PHYTOS, contributions des variables (espèces) aux nœuds de la hiérarchie du lien moyen.*

Voyons maintenant l'application des calculs de contributions à la hiérarchie ascendante de la distance moyenne, calculée sur l'indice de distance de Jaccard. Comme dans le cas quantitatif, seuls les niveaux supérieurs de l'arbre sont intéressants (Cf figure 2, construite d'après les résultats du chapitre 4) :

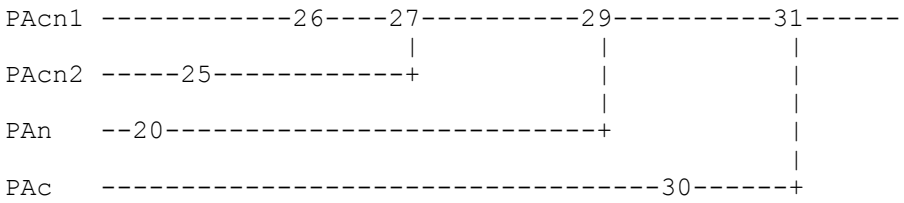


Figure 2.- *Partie supérieure de la hiérarchie de la distance moyenne. Données PHYTOS, distance de Jaccard.*

Dans le tableau 5 la disposition est que les lignes représentent les variables (espèces) tandis que les colonnes sont les nœuds successifs de la hiérarchie. Pour le nœud 31 (dernière colonne) se détachent les espèces suivantes (la contribution au nœud est entre parenthèses) :

```
117  Potentilla aurea L. (100)
144  Sempervivum arachnoideum (76)
 72  Geum montanum (73)
```

On a déjà vu, en effet, que l'espèce 117 est absente du groupe PAc (nœud 30) alors qu'elle est présente partout ailleurs (nœud 29). A l'inverse l'espèce 144 est quasi exclusive du groupe PAc tandis que l'espèce 72 en est presque totalement absente. Sautant le nœud 30 qui subdivise le groupe PAc, on peut analyser de la même façon le nœud numéro 29, pour lequel ressortent les espèces :

```
 11  Antemnaria divica (100)
 69  Gentiana nivalis (66)
 82  Homogyna alpina (L) Cass. (66)
 95  Luzula sysicata (L) DC (66)
129  Sapina glabra (Willd) Fewyl (66)
159  Trifolium pratense L (66)
165  Veronica alionci Vill. (66)
```

L'espèce 11 est totalement absente du groupe PAn (nœud 20) alors qu'elle est dans tous les relevés des groupes PAcn1 et PAcn2 qui forment le nœud 27. Les espèces 82, 95, 129, 159 caractérisent le groupe PAn bien qu'elles n'en soient pas exclusives. L'espèce 165 est absente de ce groupe, et présente dans la plupart des relevés du nœud 27.

Enfin pour le nœud 27 qui sépare les deux groupes PAcnl (nœud 26) et PAcn2 (nœud 25), on relève les espèces :

```
 55 Elyna sp. (100)
151 Salix herbacea (100)
```

4. Procédures de calculs

Pour les contributions aux noeuds d'une hiérarchie les procédures *CTRHqual* et *CTRHquan* sont disponibles, la première s'applique à des variables qualitatives et la seconde aux variables quantitatives. Pour les contributions aux classes d'une partition on a de la même façon les procédures *CTRPqual* et *CTRPquan*.

Chapitre 9

Pratique de la classification

Sans chercher à être exhaustifs nous avons jusqu'à présent examiné les méthodes typologiques les plus courantes. En étudiant leurs principes et leurs propriétés, on a noté, au passage, que chacune d'elles possède souvent plusieurs variantes ... L'utilisateur novice est donc confronté à un choix difficile qui doit être subordonné à la nature des données et à l'objectif qu'il poursuit. C'est ce qu'on examinera au paragraphe 1. En outre il est possible d'utiliser successivement deux algorithmes, l'un affinant les résultats de l'autre. De telles stratégies seront envisagées au paragraphe 2. Enfin quelques règles élémentaires d'interprétation des résultats seront établies au paragraphe 3, et deux algorithmes auxiliaires seront décrits au paragraphe 4.

1.- Choix d'un algorithme

Le choix est à faire entre quatre grandes méthodes hiérarchiques ascendantes (trois agrégations élémentaires et l'agrégation suivant le moment d'ordre deux), une méthode hiérarchique descendante et une méthode non-hiérarchique, dite agrégation autour de centres mobiles. Dans la suite on désignera ces algorithmes par leur nom générique : CAH pour les constructions ascendantes, CDH pour la construction descendante et CENMOB pour la partition par agrégation autour de centres mobiles.

1.1.- Dimensions des données

Le lecteur aura déjà remarqué que certains algorithmes nécessitent une taille de mémoire centrale plus importante que d'autres, contrainte qui est primordiale lorsqu'on travaille sur un micro-ordinateur ! Deux catégories d'algorithmes se distinguent aisément à ce sujet ; d'une part ceux qui manipulent des distances, les trois CAH élémentaires et CDH, d'autre part ceux qui travaillent directement sur les données brutes, la CAH du moment d'ordre 2 et CENMOB. Les premiers gèrent la matrice des distances en mémoire centrale, tandis que les seconds travaillent sur le tableau des données brutes.

L'avantage va, en général aux seconds. En effet supposons que l'on ait un tableau de données ayant 200 individus et 15 variables, ce qui est une disposition assez commune. Le tableau des données occupera donc 200*15 = 3000 cases, tandis que le tableau des distances nécessiterait (200*199)/2 = 19900 cases. En revanche si le nombre des variables est élevé alors les algorithmes travaillant sur les distances sont supérieurs. Dans la version actuelle des procédures Excel (Juin 2006) la programmation est faite de manière à pouvoir occuper toute la mémoire vive disponible. Toutefois il est bon de savoir qu'il y a des limites à la dimension des tableaux que l'on peut traiter.

En fait les programmes du type "centres mobiles" pourraient accepter des dimensions encore plus grandes avec des modifications mineures. Il suffirait de ne pas stocker en mémoire le tableau des données, mais de le relire à chaque fois qu'on en a besoin, les individus étant balayés séquentiellement dans les deux cas où cela se produit. Cependant un allongement considérable du temps de calcul serait à prévoir, du à la lenteur des accès disques.

1.2.- Nature des données

Lorsque les données sont quantitatives chacun des programmes peut être utilisé. Cependant il convient de refléchir si l'on doit effectuer une normalisation préalable des variables.

En revanche, avec les données qualitatives, il est obligatoire de choisir une formule de distances adaptée (voir chapitre 3). Dans ce cas il faut utiliser l'un des programmes *DisJac* ou *DisKi2*, ou tout autre programme destiné au cas de variables qualitatives, puis appliquer CAH ou CDH aux distances ainsi calculées.

1.3.- Qualité des résultats

En dehors des contraintes évoquées ci-dessus, les quatre principaux algorithmes étudiés ne donnent pas des résultats d'égale qualité. On a déjà critiqué chacun d'eux en son temps, mais il est bon de rappeler que, dans le cas où ils sont tous applicables, l'expérience permet d'établir entre eux la hiérarchie suivante (en allant du médiocre au très bon) :

CenMob --> CDH --> CAHmom2 --> CAHLM

Bien entendu ce rangement correspond aux cas les plus fréquents ; il peut arriver que, pour un exemple particulier, *CDH* ou *CAHmom2* donne une hiérarchie "meilleure" que *CAHLM*, c'est à dire plus conforme à ce qu'un examen attentif de la matrice des distances permet d'espérer. Mais dans le cas général on sera presque toujours satisfait de la hiérarchie obtenue par agrégation suivant la distance moyenne. La hiérarchie fournie par *CAHmom2* est presque toujours très voisine, dans sa structure, de celle que donne *CAHLM* bien que les niveaux d'agrégations soient fort différents.

L'ordre de préférence ci-dessus n'implique pas qu'il faille éliminer la méthode d'agrégation autour de centres mobiles. On a vu, en effet (paragraphe 1.1), que, pour certaines tailles de données, c'est la seule applicable. D'autre part en essayant un nombre suffisant de partitions initiales différentes, on parvient à des solutions satisfaisantes, indiquées par une valeur élevée du moment inter-classe.

1.4.- Temps de calcul

Du point de vue du temps de calcul la palme revient sans conteste au programme CAHMOM. Ceci est du à l'emploi de l'algorithme spécial, dit "des voisins réciproques". Bien entendu cette méthode particulière pourrait être également utilisée pour construire les hiérarchies élémentaires, mais sa programmation serait un peu plus complexe (*cf* De Rahm 1980).

2.- Stratégies

Suivant la nature des données et l'objectif à atteindre on peut utiliser une des stratégies suivantes

- Classification hiérarchique, tronquée pour donner une partition, servant de point de départ à une agrégation autour de centres mobiles
- Agrégation autour de centres mobiles pour obtenir un ensemble de classes dont les centres de gravité sont alors utilisés comme données pour une classification hiérarchique.
- D'autre part une stratégie mixte, employant conjointement analyse factorielle et classification donne souvent d'excellents résultats.

2.1.- Construction hiérarchique suivie des centres mobiles

L'objectif d'une telle stratégie est d'obtenir une partition de bonne qualité. On a vu (chapitre 5) que, quelle que soit la partition de départ, l'algorithme des centres mobiles ne peut qu'améliorer la valeur du moment inter-classe. Il est donc tentant de fournir à cet algorithme une partition initiale élaborée, au lieu de la tirer au hasard. Cela peut être fait par l'application préalable d'une CAH ou d'une CDH. En effet en "coupant" l'arbre obtenu à un endroit où la succession des niveaux d'agrégation présente un saut important, on obtient généralement une partition cohérente. On peut même, si cela semble

justifié, modifier manuellement cette partition, avant de l'introduire dans un programme d'agrégations autour de centres mobiles. L'examen soigneux de l'arbre hiérarchique est nécessaire ; on pourra éventuellement essayer plusieurs variations autour de la partition obtenue par troncature.

2.2.- Centres mobiles suivis d'une construction hiérarchique

On a remarqué ci-dessus (paragraphe 1), que, lorsque les dimensions du tableau des données sont importantes (plusieurs centaines d'objets, voire plusieurs milliers), il arrive que la seule méthode possible soit celle des agrégations autour de centres mobiles. En une telle circonstance il n'est guère possible d'essayer plusieurs partitions initiales tirées au hasard, car, à chaque tirage, le temps de calcul nécessaire au déroulement de l'algorithme peut être assez long, et l'on est donc contraint de se limiter à quelques essais, voire à un seul. La partition ainsi obtenue peut être de qualité médiocre et ne donne pas d'assurances sur la validité du nombre de classes choisi.

Plutôt que de chercher directement le regroupement des objets en un nombre restreint de classes, nous proposons, dans une étape préliminaire, d'obtenir une partition en un grand nombre de classes. Puis de prendre les centres de gravité (ou points moyens) de ces classes comme objets pour une classification hiérarchique. Supposons que l'on ait, par exemple, 5000 observations à classer. On pourra demander aux centres mobiles de créer, disons, 100 classes. Chacune d'entre elles contiendra, en moyenne, 50 observations, qui seront représentées par les valeurs moyennes de leurs variables. Ces points moyens seront alors agrégés en un temps raisonnable par une construction ascendante hiérarchique.

2.3.- Données hétérogènes, emploi de l'analyse factorielle préalable

On a déjà eu l'occasion d'examiner les avantages du pré-traitement par l'analyse factorielle (cf chapitre 3, paragraphe 1.2) mais il nous parait souhaitable de rappeler ici l'un d'eux, on relation avec la nature des données. En effet lorsque celles-ci comprennent un mélange de variables quantitatives et qualitatives, le plus simple est de rendre qualitatives celles qui ne le sont pas, par l'établissement de classes de valeurs. On considère ensuite chaque classe de valeur, ou chaque catégorie pour les variables qualitatives, comme une variable en 0 ou 1, suivant que les objets tombent dans cette catégorie ou non. Cela s'appelle mettre les variables sous forme disjonctive complète.

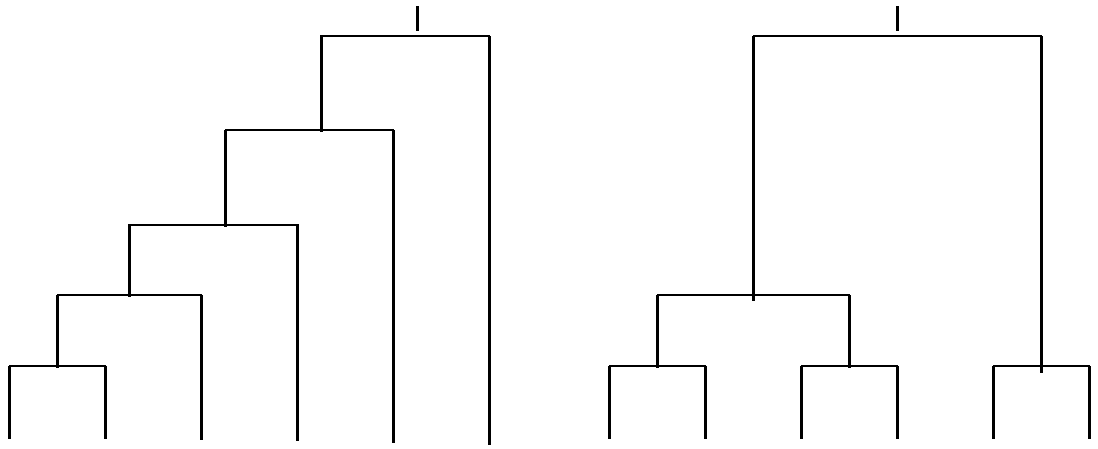
Ceci fait on peut alors calculer un indice de distance adapté à ce type de données (Distance du Khi2 ou indice de Jaccard, par exemple) ; cependant ces indices eux-mêmes existent en grand nombre et un nouveau choix délicat est donc nécessaire. Dans l'ignorance d'une bonne formule, la distance du Khi-deux donnera généralement des résultats satisfaisants. Mais, compte tenu des avantages de la méthode et de l'intérêt des résultats intermédiaires qu'elle fournit, l'emploi préalable de l'Analyse factorielle des correspondances nous parait s'imposer en de telles circonstances. On calculera ensuite la distance euclidienne usuelle sur les premiers axes retenus ; du point de vue des résultats, cette stratégie est quasi-équivalente, comme on l'a vu avec l'exemple PSYSOC, à effectuer la classification après calcul de la distance du Khi-deux sur les données brutes.

Cette stratégie apporte en outre un avantage supplémentaire. Elle transforme un grand nombre de variables qualitatives en un petit nombre d'axes factoriels, qui peuvent être considérés comme des variables quantitatives, et, par suite, les quatre grands types d'algorithmes sont applicables.

3.- Interprétation des résultats

On a vu, au chapitre 8, qu'un certain nombre de calculs supplémentaires facilitent l'interprétation des résultats, mais nous voulons parler ici d'un autre problème. Il s'agit du fait que, quelles que soient les données, les algorithmes de classification fournissent toujours une typologie. On conçoit, pourtant, que certains échantillons très homogènes, pouvant être considérés comme issus d'une

population unique, ne devraient pas donner lieu à une taxinomie. Dans le cas d'une classification hiérarchique, quelques règles permettent, *a posteriori*, d'estimer la "classifiabilité" des données.



Figures 1a et 1b . *- Les deux formes d'arbres extrêmes en classification hiérarchique.*

Deux formes d'arbres extrêmes peuvent se présenter qui sont schématisées dans les figures 1a et 1b. ans le premier cas les deux objets les plus proches constituent le noyau auquel viennent se raccrocher progressivement tous les autres objets. Dans le second cas, au contraire, se distinguent clairement des groupes bien individualisés, reliés à des niveaux élevés par rapport aux distances intra-groupes.

L'intuition suggère que, dans le premier cas, les données ne sont pas "classifiables" , tandis qu'elles le sont dans le deuxième cas. L'expérience confirme cette appréciation mais elle doit être modulée en fonction de l'algorithme utilisé pour la construction hiérarchique. En effet certains algorithmes ont une propension à donner un arbre du type 1 d'autres à donner des groupes très visibles comme dans le cas 2.

Ainsi dans l'agrégation par le saut minimum, ou lien simple, l'effet de chaîne caractéristique de cette méthode (voir chapitre 4, paragraphe 1.2), se traduit par un arbre du type 1 ; mais même en l'absence de réelle disposition en chaîne, cette méthode tend à rétrécir les intervalles de variation des distances, donc à rapprocher les niveaux d'agrégation. On n'en conclura donc pas à la non-validité des groupes observés.

La hiérarchie du diamètre, ou lien complet, présente la particularité inverse ; c'est à dire qu'elle montre des groupes bien marqués, là où, parfois, on n'a qu'une seule suite de points à faible distance les uns des autres. En bref, elle "casse" les chaînes d'objets. Bien entendu l'agrégation par la distance moyenne réalise un compromis intéressant entre les deux méthodes précédentes.

Enfin la hiérarchie du moment d'ordre deux présente le même défaut que celle du diamètre en plus poussé. Non seulement elle a tendance à fabriquer des "boules" de diamètres comparables, mais l'impression de netteté des groupes formés est encore accentuée par le fait que les niveaux de liaison ne sont pas des distances mais plutôt des carrés des distances.

4.- Un programme supplémentaires utile : troncature d'une partition

La stratégie décrite ci-dessus au paragraphe 2.1 nécessite la troncature d'une hiérarchie pour obtenir une partition de départ à introduire dans la procédure *CenMob1* d'agrégations autour de centres mobiles. Dès qu'on dépasse quelques dizaines d'objets effectuer cela à la main est fastidieux et source d'erreurs. C'est pourquoi nous avons mis au point la procédure *Troncat* pour faire cette

opération. Elle crée une partition de n objets en k classes, k étant fixé par l'utilisateur, à partir d'une hiérarchie, décrite par "ainés et benjamins" (voir chapitre 4, paragraphe 3.1). La partition obtenue est constituée par une suite de n valeurs comprises entre 1 et k. La i-ème valeur donne le numéro de la classe de l'individu i.

Chapitre 10

Conclusion

Etant données la faiblesse des connaissances mathématiques, en matière de classification, et l'impossibilité où l'on se trouve d'examiner toutes les solutions possibles, les conseils que l'on peut donner en conclusion, ne sont basés que sur des appréciations expérimentales. Rappelons que, comme tout processus d'Analyse des données, l'obtention d'une classification, se fait en trois phases principales au cours desquelles l'utilisateur est amené à faire des choix cruciaux :

- préparation des données
- traitement
- interprétation des résultats

De plus il arrive souvent que l'interprétation fasse apparaître des redondances de variables ou l'hétérogénéité de l'échantillon, ce qui amène à modifier le tableau initial et à réitérer le processus complet. Bien que chacune de ces phases, que nous allons réexaminer plus loin, pose ses problèmes propres, elles ne peuvent être totalement dissociées, les unes des autres, et elles doivent tenir compte de l'objectif global. Celui-ci peut, selon nous, être de deux sortes. Ou bien le but est d'obtenir une taxinomie de qualité, ou bien la classification n'est qu'une étape préliminaire, destinée à réduire la taille des données ou à trouver des sous-échantillons homogènes, en vue de l'application d'une autre méthode statistique (analyse factorielle, régression multiple ...). Rappelons les principaux problèmes qui se posent et comment on peut les résoudre en fonction de ces objectifs globaux.

1.- Taxinomie de qualité

Si l'on recherche avant tout la qualité des résultats on s'orientera plutôt vers la classification ascendante hiérarchique, qu'on pourra éventuellement améliorer, après troncature, par une agrégation autour de centres mobiles (cf chapitre 9).

1.1.- Préparation des données

La préparation consistera essentiellement à calculer une distance entre les objets à classer. La formule à retenir dépendra de la nature des données, qualitatives, quantitatives ou mixtes. Dans le cas purement qualitatif ou purement quantitatif diverses formules sont disponibles (voir chapitre 3 et annexe 1). On veillera à éviter les redondances : introduire deux variables mesurant le même phénomène, ou une variable dont la valeur s'obtient à partir des valeurs de deux autres variables ... Mais le mélange de variables qualitatives et quantitatives pose quelques difficultés. On est dans ce cas obligé de créer de nouvelles variables qualitatives correspondant aux classes de valeurs que l'on aura eu soin de faire pour chacune des variables quantitatives. Dans tous les cas une analyse factorielle préalable fournit généralement une base de départ solide pour la classification.

1.2.- Traitement

Les nombreuses variantes de la construction ascendante hiérarchique ne doivent pas impressionner l'utilisateur. Dans la plupart des cas l'agrégation par la distance moyenne, ou bien celle du moment d'ordre deux, lui fourniront de bons résultats (chapitres 4 et 6).

1.3.- Interprétation des résultats

Toutes les variables jouent un rôle équivalent dans la détermination des groupes d'objets, et il est rare qu'un groupe puisse être caractérisé par une plage de variation déterminée d'une seule variable. Cependant les aides à l'interprétation (chapitre 8) peuvent mettre en avant quelques unes des variables, avec des valeurs typiques pour certains groupes. On se souviendra aussi que l'ordre "horizontal " dans lequel on place les objets, en bas d'un arbre hiérarchique, est assez arbitraire puisqu'on peut faire "pivoter" un groupe sur lui-même, autour de son noeud. Autrement dit la proximité horizontale ne veut rien dire, seuls les niveaux de liaison sont à prendre en compte et ceux-ci indiquent généralement des distances moyennes entre les groupes, non entre les individus (sauf aux niveaux inférieurs).

2.- Classification en tant que pré-traitement

Le traitement de grands ensembles de données, dans le but de réduire leur taille, pose plutôt moins de problèmes. Il exclut toutes les méthodes nécessitant la gestion de la matrice des distances en mémoire centrale. Il ne reste donc que la classification ascendante hiérarchique du moment d'ordre deux, programmée selon la méthode des voisins réciproques (chapitre 6), ou bien l'agrégation autour de centres mobiles (chapitre 5). Toutefois un arbre hiérarchique portant sur des milliers d'individus est difficile à examiner et à interpréter. La méthode de choix est donc l'agrégation autour de centres mobiles.

2.1.- Préparation des données

Les deux méthodes envisageables traitent exclusivement des données quantitatives. Si l'on a des données qualitatives on devra donc obligatoirement passer par l'intermédiaire de l'Analyse factorielle des correspondances sur le tableau des données transformées en 0-1 ; cette analyse réalise en effet une sorte de "quantification" des données sur les axes factoriels.

2.2.- Traitement

Le choix de l'une ou l'autre des deux méthodes possibles tiendra compte surtout de leur fonctionnement car les contraintes liées à la taille des données sont à peu près les mêmes pour les deux méthodes. Celle du moment d'ordre deux d'une partition fournit une hiérarchie, qui devra donc être tronquée si l'on veut une partition des objets. L'agrégation autour de centres mobiles fournit directement une partition, mais elle nécessite le choix d'une partition initiale (qui peut être tirée au hasard) dont dépend le résultat final.

2.3.- Interprétation

Lorsqu'on utilise la classification comme une étape préliminaire à d'autres traitements on ne cherche pas d'interprétation aux résultats. Cependant les aides à l'interprétation sont parfois utiles pour critiquer les données avant d'aller plus avant dans leur analyse.

En résumé, on ne devra pas s'effrayer devant la variété des algorithmes possibles car le choix se limite, de fait, à deux ou trois d'entre eux pour leur qualité, ou pour leur efficacité sur de grands tableaux. D'ailleurs, lorsqu'on peut les comparer, on constate généralement un bon accord entre les résultats des différentes méthodes.

Pour des applications répétitives dans un domaine précis, l'utilisateur devra vraisemblablement faire des essais comparatifs, et choisir l'algorithme qui lui parait le mieux adapté à son problème. Cependant nous déconseillons l'adjonction de variantes personnelles qui ont trop souvent pour but de fournir des résultats en accord avec l'hypothèse que l'on veut démontrer ...

La multiplicité des algorithmes de classification ne doit pas faire oublier la multiplicité encore plus grande des traitements préliminaires des données, souvent indispensables (voir chapitre 2), et qui sont généralement décisifs pour la qualité des résultats (cf Benzécri J.P. 1973, Benzécri J.P. et F. 1980).

ANNEXE 1

Les indices de distances

N.B. Dans ce qui suit on utilise les signes mathématiques classiques suivants :

$\forall$ = pour tout … ou quel que soit …
$\in$ = appartenant à …
$\Rightarrow$ = implique
$\Sigma_k\ x_k$ = somme de tous les termes analogues x_1, x_2, etc ... en faisant varier l'indice k. Cette somme s'écrit aussi : $\Sigma\ \{x_k \mid k = 1,2,...,n\}$

1.- Généralités

Intuitivement, un indice de distance d est une formule qui permet de mesurer de combien diffèrent deux des objets que l'on étudie. C'est une évaluation de leur dissemblance ; mathématiquement, si I est l'ensemble de ces objets, d est une application (fonction) de I x I dans l'ensemble R des nombres réels positifs ou nuls, dont on exige :

1) $\forall\ i \in I,\quad d(i, i) = 0$
2) $\forall\ i, i' \in I\quad d(i, i') = d(i', i)$

Si de plus on a (inégalité triangulaire) :

3) $\forall\ i, i', i'' \in I\quad d(i,i'') \leq d(i,i') + d(i',i'')$

alors d est une véritable distance, mais cette dernière condition n'est pas indispensable pour la bonne marche des procédures de classification usuelles.

D'autre part, certains auteurs préfèrent parler en termes de ressemblance et utilisent, à cette fin, un indice de similitude s ("similarity index"), qui devra satisfaire des conditions analogues à celles de d :

1) $\forall\ i \in I,\quad s(i,i) = s_{max}$
2) $\forall\ i, i' \in I,\ s(i,i') = s(i',i)$

s_{max} est la valeur maximum que peut prendre s :

$s_{max} = Sup\ \{s(i,i') \mid i \in I,\ i' \in I\}$

elle dépend de la formule retenue, vaut généralement 1 mais peut être parfois infiniment grande.

Supposons s défini, sur I x I. Si pour tout i, et tout i' de I on pose

$d(i, i') = s_{max} - s(i, i')$

alors d sera un indice de distance. Dans ce cas, se donner l'un ou l'autre des types de mesure est équivalent, puisqu'on passe facilement de l'un à l'autre.

Remarque 1 : Certains auteurs n'imposent pas la condition 1 ; c'est à dire que l'on peut avoir $s(i, i) < s_{max}$, ainsi que $s(i, i) \neq s(j, j)$ si $i \neq j$.

Définition (Sera reprise à l'annexe 2.) : Une ordonnance sur I est une relation de préordre sur IxI, que l'on notera ≤. On aura donc :

1) réflexivité : $\forall$ i, i'$\in$I, (i, i') ≤ (i, i') ;
2) transitivité : (i, i') ≤ (j, j') et (j, j') ≤ (k, k') ⇒ (i, i') ≤ (k, k')

Remarque 2 : Ce préordre peut être non total, c'est à dire que certaines paires peuvent ne pas être comparables à certaines autres.

Remarque 3 : S'agissant d'un préordre, on n'a pas nécessairement :

(i,i') ≤ (j,j') et (j,j') ≤ (i,i') => (j,j') = (i,i')

Remarque 4 : Un indice de distance d sur I x I induit une ordonnance de la facon suivante :

(i,i') ≤ (j,j') si, et seulement si, d(i,i') ≤ d(j,j')

Un tel préordre qui est alors total, éclaire la remarque 3 : deux paires d'objets peuvent présenter le même niveau de dissemblance sans pour cela être identiques.

Nous insistons sur cette notion d'ordonnance car nous avons constaté empiriquement, et R.N. Shepard (1962) a montré, que sa seule connaissance suffit généralement pour reconstruire le nuage donné, à une homothétie près, avec une approximation d'autant meilleure que la dimension réelle du phénomène étudié est petite relativement au nombre d'observations. Autrement dit, deux nuages de points ayant des ordonnances voisines, auront des structures analogues, même si les valeurs respectives des distances sont assez différentes.

2.- Cas des données binaires

Soient i et i' deux objets quelconques de I ; ils sont représentés par deux vecteurs booléens, à n composantes si n est le nombre total d'attributs possibles.

$i = (x_1, x_2, \ldots, x_n)$ $\quad i' = (x'_1, x'_2, \ldots, x'_n)$

Pour tout k, x_k (respectivement x'_k) ne peut valoir que 0 ou 1, suivant que le caractère k est présent ou absent chez l'individu i (respectivement i'). Dans la suite, nous utiliserons les nombres suivants :

$p = \Sigma \{x_k \mid k = 1,2,\ldots,n\}$
$q = \Sigma \{x'_k \mid k = 1,2,\ldots,n\}$

p (respectivement q) est donc le nombre d'attributs possédés par i (respectivement i'). Nous appellerons c le nombre d'attributs possédés en commun par i et i', ce qui peut s'écrire :

$c = \Sigma \{ x_k x'_k \mid k = 1, 2, \ldots, n \}$

On remarque que ces quantités suffisent à exprimer le nombre d de caractères absents simultanément :

$d = n+c - (p+q) = \Sigma \{(1-x_k)(1-x'_k) \mid k = 1, 2, \ldots, n\}$

En résumé, on a la table suivante :

i\i'	1	0
1	c	p – c
0	q – c	d = n + c – p - q

où chaque case désigne le nombre d'attributs qui sont dans l'état indiqué en tête de la ligne et de la colonne correspondantes.

Nous allons maintenant énumérer, ci-dessous, les différentes formules connues comme indices de ressemblance en appelant chacune d'elles du nom du premier auteur l'ayant employée, à notre connaissance. Elles seront exposées suivant un ordre analogue à celui qui est adopté par Sokal et Sneath (1963), c'est à dire en présentant d'abord les formules où la ressemblance n'est prise en compte que par les présences communes d'attributs, puis celles où la ressemblance est comptée à la fois par les présences communes et les absences communes.

Faire un choix entre ces formules en vue d'une application précise est une tâche assez délicate, c'est pourquoi nous complèterons cet exposé de divers renseignements : intervalle de variation absolu (v.a.), c'est à dire en supposant que tous les caractères puissent prendre chacune des deux valeurs 0 ou 1, puis variation relative (v.r.) en supposant que les nombres d'attributs p et q sont fixés.

Enfin, nous nous intéresserons à la "valeur moyenne" de chacun des indices considérés. Plus précisément, on supposera que tous les caractères retenus pour la composition du tableau de données sont équiprobables, que p et q sont fixés et que l'on tire toute paire d'attributs indépendamment l'un de l'autre. Dans ces conditions, il y a p/n chances pour que i possède l'attribut k ; de même il y a q/n chances pour que i' possède k ; les deux tirages étant indépendants, il y a pq/n^2 chances pour que i et i' possèdent k ensemble, l'espérance mathématique (e.m.) de c (nombre d'attributs en commun) est donc pq/n.

Voici donc ces formules assorties, le cas échéant, de remarques ou de critiques ; elles sont toutes présentées sous forme d'indices de similitude.

2.1.- Indices où la présence des attributs joue un role prépondérant

Le souci majeur des auteurs de ces formules a été, comme on le voit sur le tableau 1, de pondérer le nombre c d'attributs communs, par les poids des deux objets considérés, c'est à dire les nombres totaux d'attributs possédés par l'un et par l'autre. Les numéros figurant dans la colonne "Note" de ce tableau 1 renvoient aux remarques ci-dessous. La colonne « Moyenne » est calculée comme l'espérance mathématique dans les conditions suivantes : les nombres p et q d'attributs des deux objets sont fixés, tous les attributs ont même probabilité d'apparition et ils sont indépendants.

N	Auteur	Formule	Etendue	Moyenne	Note
1	Russel & Rao 1940	c/n	(0,1)	pq/n	1
2	Jaccard 1908	c/(p + q - c)	(0,1)	pq/(n(p+q)-pq)	3
3	Dice 1945	2c/(p + q)	(0,1)	2pq/(n(p+q))	2, 3
4	Sokal & Sneath-2 1963	c/(2(p + q) - 3c)	(0,1)	pq/(2n(p+q)-3pq)	3
5	Kulczinski-1 1927	c/(p + q - 2c)	(0,Infini	pq/(n(p+q)-2pq)	3
6	Kulczinski-2 1927	(c/p + c/q)/2	(0,1)	(p+q)/2n	2
7	Ochiai 1957	c/Rac(p,q)	(0,1)	Rac(pq)/n	2
8	Simpson 1960	c/Min(p,q)	(0,1)	Max(p,q)/n	4
9	Kochen & Wong 1962	nc/pq	(0,n)	1	5

Tableau 1. *Indices où la présence des attributs joue un rôle prépondérant.*
p = nombre d'attributs du 1-er objet ; q = nombre d'attributs du 2-eme objet ; c = nombre d'attributs communs aux 2 objets ; n = nombre total d'attributs possibles ; Rac = racine carrée ; Min = minimum ; Max = maximum

Note 1 : Dans l'indice de Russel et Rao (numéro 1), si p=q, alors s(i,i') = p/n, s(i',i') = q/n et i ne ressemble pas à lui-même avec la même "intensité" que ne le fait i' envers lui-même.

Note 2 : Les indices de Dice (numéro 3), Kulczinski-2 (numéro 6) et Ochiai (numéro 7) ne sont autres que c divisé par la moyenne arithmétique de p et q, leur moyenne harmonique et leur moyenne géométrique, respectivement. On peut donc s'attendre à ce que les valeurs de ces indices soient voisines, s'écartant le plus les unes des autres lorsque p et q sont les plus différents (Cf. Roux G. et Roux M. 1967).

Note 3 : Les indices de Jaccard (numéro 2), Dice (numéro 3), Sokal et Sneath-2 (numéro 4) et Kulczinski-1 (numéro 5) donnent la même ordonnance. (Cf. définition de ce terme au paragraphe précédent.) Cela tient à ce qu'ils sont, tous quatre, fonctions décroissantes de (p+q)/c. L'indice de Jaccard, par exemple, peut s'écrire sous la forme s = 1 / ((p+q) / c - 1) ; on vérifiera que les trois autres indices cités se mettent sous des formes analogues. Rappelons que la structure de l'arbre, dans certaines classifications hiérarchiques, ne dépend que de l'ordonnance, elles donnent donc les mêmes résultats avec ces quatre indices (voir chapitre 4, paragraphe 1.2).

Note 4 : Dans l'indice de Simpson (numéro 8) comme dans tous les autres, c a pour valeur minimum soit zéro, si p+q < n, soit (p+q-n) / Min (p,q). Dans le premier cas, qui est fréquent dans de nombreuses disciplines comme l'écologie végétale ou animale, l'archéologie, etc ... l'intervalle de variation, lorsque p et q sont fixés, est [0, 1]. Il est donc indépendant de p et q, ce qui n'est pas le cas pour les autres indices, en général.

Note 5: Pour l'indice de Kochen et Wong (numéro 9) l'espérance mathématique (dans les conditions décrites au début de ce paragraphe) est constante, mais les objets de faible poids sont avantagés.

2.2.- Indices où les présences et absences d'attributs jouent des rôles équivalents

Le titre de ce paragraphe est un peu abusif car on sait que c et d ne sont pas indépendants (voir début paragraphe 2), il ne s'agit donc que d'une symétrie d'écriture. La plupart de ces indices, décrits dans le tableau 2 s'obtiennent à partir de leur homologue (colonne H) du tableau précédent où d est introduit de facon naturelle.

Compte tenu que la valeur moyenne de d est égale à (n-p)(n-q)/n, on en déduit facilement les valeurs moyennes de ces indices. Nous signalerons dans la Note numéro 10 les valeurs remarquables de ces moyennes. Voici quelques commentaires sur ces indices.

Note 6 : Les trois premiers indices du tableau (numéros 11, 12 et 13) donnent la mème ordonnance car ils sont tous trois fonctions décroissantes de n/(c+d).

Note 7 : Nous avons construit les indices numéro 17 et 18 par analogie avec les formules numéro 8 et 9 du tableau 1.

Note 8 : La valeur maximum, n, de l'indice numéro 18 est atteinte pour c = d = p = q = n-1 ; on suppose en effet, que tout objet possède au moins un attribut, et au plus n-1.

Note 9 : Si s' est l'indice de Sokal et Michener (numéro 11) et si s désigne le coefficient numéro 19, alors on a : s = 2s' - 1. Les propriétés de s se déduisent donc facilement de celles de s', outre que ces deux coefficients ont mème ordonnance.

No	Auteurs	Formule	Etendue
11	Sokal & Michener 1958	(c + d)/n	[0, 1]
12	Sokal & Sneath-1 1963	2(c + d)/(n + c + d)	[0, 1]
13	Rogers & Tanimoto 1960	(c + d)/(2n - (c + d))	[0, 1]
14	Sokal & Sneath-3 1963	(c + d)/(p + q - 2c)	[0, Infini]
15	Sokal & Sneath-4 1963	S1 = c/p + c/q S2 = d/(n - p) + d/(n - q) s = (S1 + S2)/4	[0, 1]
16	Sokal & Sneath-5 1963	cd/Rac(pq(n-p)(n-q))	[0, 1]
17	Roux-1 1985	D1 = Min(p, q) D2 = Min(n - p, n - q) s = (c + d)/(D1 + D2)	[0, 1]
18	Roux-2 1985	(n cd)/(pq(n - p)(n - q))	[0, n]
19	Hamann 1961	((c + d) - (p - c) - (q - c))/n	[-1, +1]
20	Yule 1911	N = cd - (p - c)(q - c) D = cd + (p - c)(q - c) s = N/D	[-1, +1]
21	Phi de Pearson	N = cd - (p - c)(q - c) D = Rac(pq(n - p)(n - q)) s = N/D	[-1, +1]

***Tableau 2**. Indices où les présences et les absences communes d'attributs jouent des rôles equivalents. p = nombre d'attributs du 1-er objet ; q = nombre d'attributs du 2-ème objet ; c = nombre d'attributs communs aux 2 objets ; d = nombre d'attributs absents simultanément dans les 2 objets ; n = nombre total d'attributs possibles ;* $\Sigma \{x_k \mid k = 1,2,...,n\}$
Rac = racine carrée ; Min = minimum ; Max = maximum

Remarque : le coefficient Phi (no 21) est égal au Khi-2 de contingence au coefficient 1/n près.

Note 10 : L'indice de Yule (numéro 20) possède l'intéressante propriété d'avoir un intervalle de variation s'étendant de -1 à +1 mème lorsque p et q sont fixés (cf remarque 4, paragraphe 2.1).

Note 11 : Les indices suivants ont des valeurs moyennes indépendantes de p et q (cf remarque 5, paragraphe 2.1) :

l'indice numéro 15 a pour valeur moyenne 1/2
" numéro 18 " " " 1
" numéro 20 " " " 0
" numéro 21 " " " 0

3.- Cas des données quantitatives

3.1.- Coefficients de corrélation

La plupart des coefficients de corrélation ont été créés avec l'intention de mesurer la ressemblance entre caractères. Pour évaluer la similitude entre individus ils devraient être employés avec circonspection.

Dans ce qui suit x(i, j) désigne la valeur de la j-ème variable pour l'objet i. Les formules donnent, selon l'usage, la corrélation entre variables ; il faudrait intervertir les rôles des indices i et j pour obtenir la corrélation entre observations

Coefficient de Bravais-Pearson (usuel)

$$s(j,j') = \Sigma_i \{[x(i,j)-m(j)]\,[x(i,j')-m(j')]\} / [s(j)\,s(j')]$$

m(j) et m(j') désignent les moyennes des variables j et j'
s(j) et s(j') désignent les écarts-types des variables j et j'

$$s^2(j) = \Sigma_i [x(i,j)-m(j)]^2 / n$$

Coefficient de rangs de Spearman (1904)

En supposant que, pour chaque variable j, les valeurs ont été rangées par ordre croissant, on désigne par R(i, j) le rang de l'observation i pour la variable j

$$s(j, j') = 1 - 6 \Sigma_i [R(i, j)-R(i, j')]^2 / (n(n^2 - 1))$$

Coefficient de rangs de Kendall (1938)

Dans ce coefficient, il faut, pour chaque variable j, comparer deux à deux toutes les observations. On pose :

$R_j(i, i') = 1$ si $x(i,j) > x(i',j)$
$R_j(i, i') = 0$ si $x(i,j) = x(i',j)$
$R_j(i, i') = -1$ si $x(i,j) < x(i',j)$

$$s(j, j') = 2 \Sigma_{i<i'} R_j(i, i') R_{j'}(i, i') / (n(n-1))$$

Remarque : L'avantage des coefficients de rangs est qu'ils sont indépendants de l'origine et de l'échelle des variables.

3.2.- Mesures de distances

Les formules ci-dessous expriment la distance entre deux observations i et i'. Ces formules utilisent la quantité :

$$D(j) = |x(i, j) - x(i', j)|$$

où x(i, j) est la valeur à l'intersection de la ligne i et de la colonne j du tableau rectangulaire des données (les observations sont supposées placées en lignes). D(j) est la valeur absolue de la différence des valeurs de la variable j pour les deux observations i et i'. On l'appelle parfois l'écart entre les deux observations i et i' pour le caractère j.

Ecart moyen (Czekanovski, 1932)

$$d(i,i') = \Sigma_j D(j) / p$$

p désigne le nombre de variables.

Ecart maximum

$$d(i,i') = \mathrm{Max}_j D(j)$$

Distance euclidienne usuelle

$$D^2(i, i') = \Sigma_j D^2(j)$$

Cette distance est particulièrement sensible à l'échelle choisie pour chacune des variables ; c'est pourquoi on lui préfère souvent une formule introduisant des coefficients de pondération w.

Distance euclidienne pondérée

$$d^2(i, i') = \Sigma_j w(j) D^2(j)$$

w(j) = pondération affectée à la variable j. L'usage est de prendre pour pondération l'inverse de la variance de j :

$$w(j) = 1 / s^2(j)$$

mais tout autre système de pondérations est possible, à condition que celles-ci soient positives.

Distance de Manhattan (Métrique L1)

$$d(i, i') = \Sigma_j D(j)$$

Distance de Chebychev (Métrique L-infini)

$$d(i, i') = \mathrm{Max}_j D(j)$$

Coefficient de Lance et Williams (1966)

$$d(i, i') = \Sigma_j D(j) / \Sigma_j [x(i, j) + x(i', j)]$$

C'est une généralisation du coefficient de Dice pour les données binaires (sous forme de distance).

Coefficient de divergence (Clark, 1952)

$$d^2(i,i') = (1/p)\ \Sigma_j\ D^2(j)/[x(i,j) + x(i',j)]^2$$

Ce coefficient varie entre 0 (observations identiques) et 1.

Distance du Khi-2 (Variables qualitatives ou effectifs)

Ici on change la définition de D(j) :

```
D(j)   = x(i, j)/x(i, .) - x(i', j)/x(i', .)

x(i, .) = somme des termes de la ligne i
x(., j) = somme des termes de la colonne j.
w(j)   = 1/x(., j)
```

$$d^2(i,i') = \Sigma_j\ w(j)\ D^2(j)$$

Particulièrement adaptée au cas des tableaux homogènes d'effectifs, ou de grandeurs additives (voir exemple PSYSOC, chapitre 2), la distance du Khi-2 impose une double pondération, sur les lignes et sur les colonnes du tableau des données.

4.- Conclusion

Les formules de distances, comme de similitudes, sont très nombreuses, mais il est déconseillé de choisir une formule inusitée sans raison valable. En ce qui concerne les données binaires (qualitatives) deux familles d'indices se distinguent, à l'intérieur desquelles le choix d'une formule influe peu sur le résultat de la classification. D'autres formules ont été proposées ailleurs qui font intervenir la notion de probabilité (voir Goodall 1966, Lerman 1981), ou la théorie de l'information (voir Estabrook 1967) ; mais leur complication et le faible avantage qu'elles apportent nous ont conduit à les écarter de cet inventaire.

ANNEXE 2

Hiérarchies et ultramétriques

1.- Généralités

1.1.- Hiérarchie et ordonnance

Dêfinition 1 (Benzécri, 1966) : Soit I un ensemble fini et H un ensemble de parties de J. Nous dirons que H est une hiérarchie sur I si :

1) $I \in H$
2) Pour tout $i \in I$ on a $\{i\} \in H$
3) Quels que soient h et h', éléments de H, si $h \cap h' \neq \phi$ alors on a soit $h \sqsubset h'$, soit $h' \sqsubset h$

Un couple (I,H) formé d'un ensemble fini I et d'une telle hiérarchie H peut être représenté comme un arbre dont les noeuds (traits horizontaux) symbolisent les diverses parties appartenant à H ainsi l'arbre ci-dessous correspond à la hiérarchie H formée des parties suivantes :

h1 = {1} ; h2 = {2} ; h3 = {3} ; h4 = {4} ; h5 = {5}
h6 = {2, 5} ; h7 = {4, 3}
h8 = {2, 3, 4, 5} ; h9 = {1, 2, 3, 4, 5} = I.

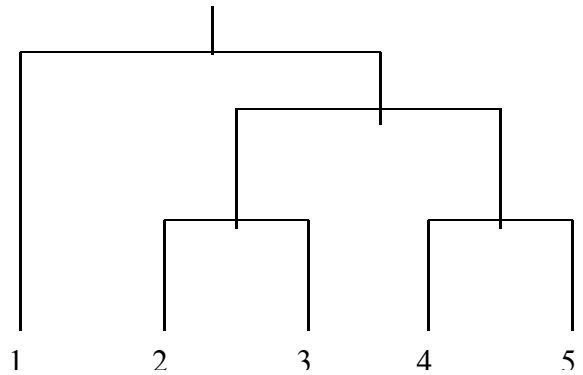


Figure 1.- *Exemple simple de hiérarchie*

Définition 2 : (Benzécri, 1966) Un ensemble I est dit muni d'une ordonnance s'il existe une relation d'ordre total sur les paires d'éléments de I.

C'est à dire que, quels que soient les éléments i, j, k, l de I, l'une ou l'autre des expressions suivantes est vraie :

(i, j) < (k, l)
(k, l) < (i, j)
(i, j) = (k, l).

Nous préférons distinguer l'égalité du cas où une paire est effectivement différente de l'autre, étant entendu que la dernière des relations ci-dessus signifie, non pas que les deux paires sont constituées des mêmes éléments, mais que les éléments qui les constituent se ressemblent autant dans la première paire que dans la seconde.

Il est évident que toute métrique d sur I induit une ordonnance en déclarant :

(i, j) < (k, l) si et seulement si d(i, j) < d(k, l) (voir annexe 1).

D'autre part une hiérarchie H sur un ensemble fini I induit une relation d'ordre (non total, en général) sur les paires d'éléments de I de la façon suivante : on dira que (i, j) < (k, l) s'il existe une partie h de H contenant i et j, telle que l'on ait :

soit l ∈ h et k ∉ h, soit l ∉ h et k ∈ h.

Si une telle partie h n'existe pas c'est que la situation est la suivante : toute partie h qui contient i et j, soit contient aussi k et l, soit ne contient ni l'un ni l'autre. Deux éventualités se présentent alors ; ou bien il existe h' ∈ H, contenant k et l mais ne contenant pas i et j, auquel cas (i, j) et (k, l) ne sont pas comparables, ou bien une telle partie h' n'existe pas et alors (i, j) < (k, l). En notation arborescente ces deux cas donnent les arbres suivants :

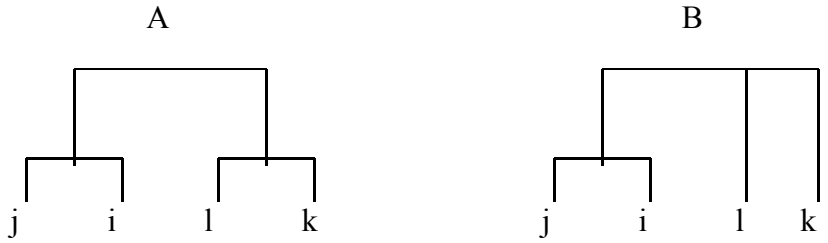


Figure 2. *Comparaison de paires d'objets*
A : Existence de h' ; B : h' n'existe pas

Enfin, si toute partie de H qui contient i et j contient aussi k et l, nous dirons que (i, j) = (k, l).

1.2.- Hiérarchie indicée et ultramétrique

Définition 1 (Benzécri 1966) : Une hiérarchie H sur un ensemble I fini est dite indicée s'il existe une application
x : H → [0, 1] telle que

1) si h ∈ H est réduite à un élément, alors x(h) = 1
2) si h ⊏ h' ∈ H alors x(h) > x(h')

On remarque immédiatement qu'une telle application permet de définir sur I un "indice de similarité" s, c'est à dire une mesure de la ressemblance, (cf Benzécri, 1966 et Roux, 1967) entre les éléments de I de la manière suivante :

Pour toute paire i, i' ∈ I, s(i, i') est le plus grand nombre x(h) tel que {i, i'} ⊏ h ∈ H et x(h) = s(i, i')

De plus on vérifie aisément que d(i, i') = 1 - s(i , i') constitue une distance sur I. Une telle hiérarchie définit donc une véritable ordonnance et non plus un ordre partiel sur les paires d'éléments de I.

Définition 2 (Bourbaki, 1958) : Une distance d sur un ensemble E est dite ultramétrique si elle vérifie, pour tout triplet de points i, j, k de I, la condition :

$$d(i,k) \leq \text{Max}\,[d(i, j), d(j, k)] \qquad (1)$$

Il est clair que la distance d, définie ci-dessus pour les indices de similarité est ultramétrique ; en effet, pour tout triplet de points de I, il ne peut y avoir que deux situations : ou bien toute partie de H qui contient deux des points, soient i et j, contient aussi le troisième, soit k, ou bien il existe h ∈ H telle que i, j ∈ h et k ∉ h. Dans le premier cas, d'après la définition

$$d(i, j) = d(j, k) = d(i, k)$$

et la relation (1) est bien vérifiée (triangle équilatéral) ; dans le second cas on a :

$$d(i,j) < d(i,k), \quad d(i,j) < d(j,k) \text{ et } d(i,k) = d(j,k)$$

d'après la définition de d, où l'on voit que (1) est encore vérifiée.

Réciproquement, à toute distance ultramétrique d sur un ensemble fini I, on peut faire correspondre une hiérarchie indicée unique H, dont s = 1 - d soit l'indice associé. En effet, la relation $s(i, i') \geq x$ (ou $d(i, i') \leq 1 - x$) est une relation d'équivalence sur I dont les classes définissent une partition P(x) unique, pour chaque x. H est alors déterminée par les parties h telles qu'il existe $x \in [0, 1]$, dont la partition P(x) contient h comme l'une de ses composantes. L'indice x(h) est alors le plus grand x tel que $h \in P(x)$.

Ces définitions et propriétés appellent quelques remarques :

Remarque 1 : La relation (1) entraîne l'inégalité triangulaire de sorte que toute application d de I x I dans R vérifiant (1), et les conditions

2) $d(i, j) = 0 \Rightarrow i = j$,
3) $\forall i, i' \in I : d(i, i') = d(i', i)$

est une distance ultramétrique.

Remarque 2 : Si d est une ultramétrique on peut démontrer que $d(i, k) \neq d(j, k)$ entraîne $d(i, j) = \text{Max}[d(i, k), d(j, k)]$, de sorte que tout triangle est isocèle avec la base inférieure aux cotés égaux. En effet, on n'enlève pas de généralité à supposer que $d(i, k) < d(j, k)$, donc $d(i, j) \leq d(j, k)$ d'après (1).

Toujours d'après (1), on a : $d(j, k) \leq \text{Max}[d(i, j), d(i, k)]$; comme $d(i, k) < d(j, k)$ par hypothèse, on a nécessairement $d(j, k) \leq d(i, j)$, donc $d(i, j) = d(j, k)$.

La correspondance entre hiérarchie indicée et ultramétrîque nous permet de poser le problème de la classification en termes plus précis que ceux de notre introduction (chapitre 1, paragraphe 2). Ce problème peut en effet être considéré comme la recherche de l'ultramétrique la plus proche de la métrique donnée. Par "proche" nous entendons ressemblante au sens d'un certain critère donné à l'avance. Malheureusement l'ensemble des métriques n'a pas la structure d'un espace vectoriel, et le sous-ensemble des ultramétriques ne peut donc pas avoir de propriété remarquable comme, par exemple, celle d'être un sous-espace ou un convexe, sous-ensembles sur lesquels on sait abaisser une perpendiculaire.

Cependant nous verrons au paragraphe 2 que, pour un critère assez fruste (Relation d'ordre) et pour une classe particulière d'ultramétriques (Ultramétriques inférieures) il existe une solution optimale.

2.- Une ultramétrique particulière : la sous-dominante

N.B. Dans ce paragraphe l'abréviation J.J.S. renvoie à l'article de Jardine C. J., Jardine N. et Sibson R.(1967).

2.1.- Relation d'ordre sur les métriques

Définition 1 (J.J.S.) : Soit un ensemble fini I, muni de deux métriques d et d'. On dira que d est inférieure à d' si, pour tout couple de points i, j ∈ I, on a : d(i, j) ≤ d'(i, j).

On vérifie facilement que c'est une relation d'ordre sur l'ensemble des métriques sur I.

Remarque 1 : Une métrique peut être inférieure à une autre et avoir la même ordonnance mais ce n'est pas toujours le cas on peut avoir une métrique inférieure à une autre mais n'ayant pas la même ordonnance et l'on peut avoir, aussi, deux métriques de même ordonnance sans qu'elles soient comparables.

Définition 2 : Soit un ensemble I fini, muni d'une famille $\{d_m \mid m \in M\}$ de métriques, indexée par M, fini ou non. Nous dirons que cette famille est bornée si pour tous i, i' ∈ I il existe b(i, i') tel que, pour tout m ∈ M, d(i, i') ≤ b(i, i').

Il en résulte immédiatement, comme I est fini et que l'ensemble des paires (i, i') est fini aussi, qu'il existe b majorant de tous les d(i, i') à savoir le Max {b(i, i') | i, i' ∈ I }.

Définition 3 : Soit un ensemble fini I, muni d'une famille bornée de métriques $\{d_m \mid m \in M\}$. Nous appellerons enveloppe supérieure de la famille, l'application de I x I dans R définie, pour tout (i, i') ∈ I x I, par :

$$(i, i') \rightarrow \text{Sup } \{d_m(i, i') \mid m \in M\}$$

Proposition : L'enveloppe supérieure d'une famille bornée d'ultramétriques sur un ensemble fini I est une ultramétrique sur I.

1) Les d_m étant des métriques, si i = i' on a pour tout m ∈ M : d_m (i, i') = 0 donc

$$\text{Sup } \{d_m(i,i') \mid m \in M\} = 0$$

Réciproquement, si l'on a : Sup $\{d_m(i,i') \mid m \in M\} = 0$, comme les d_m sont des applications positives celà entraîne que pour tout m ∈ M : $d_m(i, i') = 0$, donc i = i'.

2) On a pour tout m ∈ M : $d_m(i,i') = d_m(i',i)$ ce qui entraine :

$$\text{Sup } \{d_m(i, i') \mid m \in M\} = \text{Sup } \{d_m(i', i) \mid m \in M\}$$

3) Démontrons maintenant la relation ultramétrique (1) du paragraphe 1.2, pour l'enveloppe supérieure. La conclusion s'écrit :

$$\text{Sup } \{d_m(i,i') \mid m \in M\} \leq \text{Max } [\text{Sup } \{d_m(i,i'') \mid m \in M\}, \text{Sup } \{d_m(i',i'') \mid m \in M\}]$$

ou encore

$$\text{Sup } \{d_m(i,i') \mid m \in M\} \leq \text{Sup } \{\text{Max } [d_m(i, i''), d_m(i', i'')] \mid m \in M\}$$

avec pour hypothèse :

$$\forall m \in M, \forall i, i', i'' \in I : \quad d_m(i, i') \leq \text{Max } [d_m(i, i''), d_m(i', i'')]$$

S = Sup $\{d_m(i,i') \mid m \in M\}$ existe, car, pour tout i, i' ∈ I, $d_m(i,i')$ est borné. Cela signifie que pour tout ε > 0 il existe m* tel que S - ε < d_{m^*} (i,i').

Mais par hypothèse d_{m^*} (i,i') ≤ Max [d_{m^*} (i,i''), d_{m^*} (i' , i'')] et par passage à la borne supérieure

$$d_{m^*}(i,i') \leq \text{Sup } \{\text{Max } [d_m(i,i''), d_m(i',i'')] \mid m \in M\}$$

ce qui entraîne, car ε est quelconque, que

$$S \leq \text{Sup } \{\text{Max } [d_m(i,i'), d_m(i',i'')] \mid m \in M\}$$

Remarque 2 : On aurait pu définir d'une façon analogue l'enveloppe inférieure d'une famille de métriques, mais on n'aurait pu démontrer de proposition analogue à la précédente comme le prouve le contre-exemple suivant :

$$d_1(j, k) = 4,\ d_1(j, l) = 1,\ d_1(k, l) = 4$$
$$d_2(j, k) = 3,\ d_2(j, l) = 3,\ d_2(k, l) = 2$$

on a alors :

$$\text{Inf } [d_1(j, k), d_2(j, k)] = 3$$
$$\text{Inf } \{d_1(j, l), d_2(j, l)] = 1$$
$$\text{Inf } [d_1(k, l), d_2(k, l)] = 2$$

d'où :

$$\text{Max } [\text{Inf } \{d_1(j, l), d_2(j, l)\} , \text{Inf } \{d_1(k, l), d_2(k, l)\}] = 2$$

qui n'est pas supérieur à Inf $\{d_1(j, k), d_2(j, k)\}$ comme l'exige la relation (1) du paragraphe 1.2, ci-dessus.

2.2.- Ultramétrique "sous-dominante" d'une métrique donnée

Définition (J.J.S.) : Etant donnée une métrique δ quelconque sur un ensemble fini I, pour l'ensemble des ultramétriques inférieures à δ, celle-ci constitue un ensemble de majorants :

$$\{ \delta(i, i') \mid i \in I, i' \in I \}$$

La famille des ultramétriques inférieures à δ est donc une famille bornée. Cette famille a donc une enveloppe supérieure qui sera appelée ultramétrique sous-dominante de δ (ou plus brièvement "la sous-dominante" de δ).

Proposition : La construction ascendante hiérarchique du saut minimum fournit la sous-dominante.

(Nous reprenons ici la démonstration de Benzécri 1973). On appelle encore δ la distance initiale et d sa sous-dominante. On désigne par d_1, d_2, ..., d_k les états successifs de la distance d en cours de construction, n étant le nombre d'éléments de l'ensemble I à classer ; au début on a $d_l = \delta$. Au pas h de l'algorithme on suppose qu'on forme le groupe a par fusion des deux groupes s et s'. A chaque pas de la construction le recalcul des distances fait que les nouvelles distances sont, soit égales, soit inférieures aux distances de l'étape précédente. Par conséquent l'ultramétrique finale d est inférieure à la distance initiale.

L'ultramétrique construite est donc bien inférieure à δ. On va montrer maintenant, par récurrence, que l'ultramétrique inférieure maxima d^*, est inférieure à l'ultramétrique construite par le saut minimum.

Au début de l'algorithme $d^* \leq d_1 = \delta$. On va donc montrer que si $d^* \leq d_{h-1}$ alors $d^* \leq d_h$. Si deux points n'appartiennent ni à s ni à s' alors leur distance n'est pas modifiée par la fusion de ces deux groupes. De même si deux points appartiennent au même groupe, s ou s', leur distance est inchangée. Si $i \in s$

et si i' ∈ s' leur distance avant agrégation est la même que la distance d(s, s') entre les deux groupes, et elle est encore inchangée après agrégation.

Examinons le cas d'un point u n'appartenant ni à s , ni à s' et sa distance $d^*(u, i)$ à un point i de s . Soit i' un troisième point appartenant à s'. d^* étant ultramétrique deux cas sont alors possibles (triangles isocèles, remarque 2 ci-dessus) :

Cas 1 : $d^*(u, i) = d^*(u, i') \geq d^*(i, i')$

Par hypothèse de récurrence on a : $d^*(u, i) \leq d_{h-1}(u, i)$ et $d^*(u, i') \leq d_{h-1}(u, i')$ donc

$$d^*(u, i) \leq \text{Min}\,[d_{h-1}(u, i), d_{h-1}(u, i')] = d_h(u, i)$$

Cas 2 : $d^*(u, i) = d^*(i,i') \geq d^*(u, i')$ et le cas analogue $d^*(u, i') = d^*(i, i') \geq d^*(u, i)$

Par hypothèse de récurrence on a : $d^*(i, i') \leq d_{h-1}(i, i')$ et $d_{h-1}(i, i') = d(s, s')$. Or si on fusionne s et s' c'est parce que la distance entre ces deux groupes est la plus petite des distances intergroupes donc :

$$d(s, s') \leq d_{h-1}(u, i)$$
$$d(s, s') \leq d_{h-1}(u, i')$$

d'où :

$$d^*(u, i) = d^*(i, i') \leq \text{Min}\,[d_{h-1}(u, i), d_{h-1}(u, i')] = d_h(u, i')$$

Ainsi la propriété $d^* \leq d_k$ est vraie. Mais comme d^* est la plus grande des ultramétrîques inférieures à δ, $d_k \leq d^*$ ce qui entraîne $d_k = d^*$.

BIBLIOGRAPHIE

Anderberg M.R.(1973). Cluster analysis for applications. 359p. Academic Press, New York, London.

Benzécri J.P.(1964). Analyse factorielle des proximités. Publication de l'Institut de Statistique de l'Universîté de Paris, Paris.

Benzécri J.P.(1966). Leçons sur l'analyse factorielle et la reconnaissance des formes. Cours du 3ème cycle, ISUP, Paris.

Benzécri J.P. et coll.(1973). L'Analyse des données. Tome 1: La Taxinomie. 615p. Dunod, Paris.

Benzècri J.P.(1982). Histoire et préhistoire de l'Analyse des données. 159 p. Dunod, Paris.

Benzécri J.P. et F. Benzécri (1980). Pratique de l'Analyse des données. Analyse des correspondances, exposé élémentaire. 424p. Dunod, Paris.

Bertier P. et Bouroche J.M.(1975). Analyse des données multidimensionnelles. 270p. PUF, Paris.

Boley D. (1998). Principal directions divisive partitioning. Data mining and knowledge discovery. 2 : 325-344.

Bourbaki N.(1958) Livre III, chap. 9, Utilisation des réels en topologie, [§2, Ex. 4] Hermann,. Paris.

Bouroche J.M. et Saporta G.(1980). L'Analyse des données. 125p. Collection Que sais-je ?, PUF, Paris.

Caillez P. et Pagès J.P.(1976). Introduction à l'analyse des données. 616p. Ed. SMASH (9 rue Duban 75016 Paris), Paris.

Chandon J.L. et Pinson S.(1981). Analyse typologique. 254p. Masson, Paris.

Chavent M., Guinot C., Lechevallier Y., Tenenhaus M. (1999). Méthodes divisives de classification et segmentation non supervisée : recherche d'une typologie de la ,peau humaine saine. Rev. Stat. Appl. XLVII(4) : 87-99.

Clark P.J.(1952). An extension of the coefficient of divergence for use with multiple characters. Copeia, 2 : 61-64.

Cramer P.J.(1946). Mathematical methods of statistics. 575p. Princeton University press, Princeton.

Czekanowski J.(1932). "Coefficient of racial likeness und durchschnittliche differens". Anthrop. Anz., 9 : 227-249.

De Lagarde J.(1983). Initiation à l'analyse des données. 158p. Dunod, Paris.

De Rham C.(1980). La classification hiérarchique selon la méthode des voisins réciproques. Cah. Ana. des données, vol. V, no 2 : 135-144.

Dice L.R.(1945). Measures of the amount of ecologic association between species. Ecology 26 : 297-302.

Diday E., Lemaire J., Pouget J., Testu F.(1982). Eléments d'analyse des données. 462 p. Dunod, Paris.

Diday E.(1971). La méthode des nuées dynamiques. Rev. Stat. appliquée, vol. XIX, no 2 : 19-34.

Edwards A.W.F. and Cavalli-Sforza L.L. (1965). A method for cluster analysis. Biometrics, 21: 362-375.

Escofier B. et J. Pagès (1990). Analyses factorielles simples et multiples. 2-ème édition, Dunod, Paris, 266 p.

Estabrook G.F.(1967). An information theory model for character analysis. Taxon 16 : 86-97.

Everitt B.(1974). Cluster analysis. 122 p. Heinemann Educational Books, London.

Fages R.(1978). La notion de dispersion en classification automatique. Communication aux Journées de Statistique. Nice, 22-26 Mai 1978.

Fénelon J.P.(1981). Qu'est-ce que l'analyse des données ?. 311 p. Ed. Lefonen (26 rue des Cordelières 75013 Paris), Paris.

Foucart T.(1982). Analyse factorielle. Programmation sur micro-ordinateur. Masson, Paris.

Gondran M.(1975). Valeurs propres et vecteurs propres en classification hiérarchique. Communication aux journées d'Etude sur les Problèmes d'Analyse et d'Ajustement de tableaux statistiques, INSEE, Nantes, 23-25 Avril 1975.

Goodall D.W.(1966). A new similarity index based on probability. Biometrics : 882-907.

Guinochet M.(1955). Logique et dynamique du peuplement végétal. 144p. Masson, Paris.

Guinochet M.(1973). Phytosociologie. 227p. Masson, Paris.

Hubert L.(1973). Monotone invariant clustering procedures. Psychometrika 3O, 1.

Jaccard P.(1908). Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat., 44 : 223-270.

Jambu M. et Lebeaux M.O.(1978). Classification automatique pour l'Analyse des données. Tome 1.- Méthodes et Algorithmes (312p.), Tome 2.- Logiciels (400p.). Dunod, Paris.

Jardine N. and Sibson R.(1971). Mathematical Taxonomy. 286p. Wiley and sons, New York, London.

Jardine C. J., Jardine N., Sibson R.(1967). The structure and construction of taxonomic hierarchies. Mathematical Bioscience : 175-195.

Kendall M.G. (1938). A new measure of rank correlation. Biometrika, 30(1-2) : 81-93.

Kochen M. et Wong E.(1962). Concerning the possibility of a cooperative information exchange. IBM journal of Research and Development, 6 : 270-271.

Kulczinski S.(1927). Die Pflanzenassoziationen der Pieninen (En polonais, résumé en allemand). Bull. Intern. Acad. Pol. Sci. Lett. Cl. Sci. Math. Nat., B (Sci. Nat.), Suppl. 2 : 57-203.

Lance, G. N and W. T. Williams (1966). Computer programs for hierarchical polythetic classification. Comput. J. 9 : 60 – 64.

Lebart L., Morineau A., Fénelon J.P.(1982). Traitement des données statistiques. 518p. Dunod, Paris.

Lefebvre J.(1983). Introduction aux analyses statistiques multidimensionnelles. 275p. Masson, Paris.

Lerman I.C.(1970). Les bases de la classification automatique. 117p. Gauthier-Villars, Paris.

Lerman I.C.(1981). Classification et analyse ordinale des données. 740p. Dunod, Paris.

Reinert M. (1983). Une méthode de classification descendante hiérarchique. Cahiers analyse des donnéesd, VIII(2) : 187-198.

Roux G. et Roux M.(1967). A propos de quelques méthodes de classification en phytosociologie. Rev. Stat. Appl. vol. XIV no 2 : 50-72.

Roux M. et Guittonneau G.G.(1977). Sur la taxinomie du genre Erodium. Cah. Ana. des données, vol. II, no 1 : 97-113.

Roux M. (1985). Algorithmes de classification. 151 p., Masson, Paris.

Roux M. (1995). About divisive methods in hierarchical clustering. In "Data Science and Its Applications", Y. Escoufier, C. Hayashi, B. Fichet, N. Ohsumi, E. Diday, Y. Baba, L. Lebart (Eds) Acad. Press, Tokyo, pp 101-106.

Saporta G. (1990). Probabilités, analyse des données et statistique. Editions Technip, Paris, 493 p.

Sokal R.R. et Sneath P.H.A.(1963). Principles of Numerical Taxonomy. 359p. Freeman and co., San Francisco, London.

Shepard R.N.(1962). The analysis of proximities : scaling with an unknown distance function. I. Psychometrica, vol.27, no 2.

Spearman C. (1904). The proof and measurement of association between two things. American J. Psychology, 15 (88).

Todd E.(1979). Le fou et le prolétaire. Le Livre de Poche, Robert Laffont, Paris.

Volle M.(1978). Analyse des données. 265p. Economica, Paris.

Ward J.H.(1963). Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. 58 : 236-244.

Williams W.T and Lambert J.M.(1959). Multivariate methods in plant ecology. I. Association analysis in plant communities. J. Ecology 47 : 83-101.

INDEX

N.B. Les références indiquent successivement le numéro du chapitre et du paragraphe concernés. Ainsi "c3-1.2" désigne le paragraphe 1.2 du chapitre 3. Quand il n'y a pas de numéro de paragraphe cela signifie que tout le chapitre est consacré à la notion que l'on recherche. Les références "a1" ou "a2" désignent respectivement les annexes 1 et 2 ; enfin la lettre "b" renvoie à la bibliographie.

Agglomération (Voir agrégation)	
Agrégation(s)	
Autour de centres mobiles	c5 ; c10
Par le diamètre ou Lien complet	c4-1 ; c9-3
Par la distance moyenne	c4-1 ; c9-3
Par le lien simple ou Saut minimum	c4-1 ; c9-3
Par le moment d'ordre deux	c6 ; c9-3
Successives	c1 ; c4
Analyse factorielle	
Des correspondances	c2 ; c3-1.2
En composantes principales	c3-1.2
Prétraitement par	c3-1.2
Benzécri J.P.	c2-1 ; c2-2 ; c3-2 ; c6 ; a2-1 ; b
Bourbaki N.	a2-1 ; b
CAHLM	c4-3 ; c9-1
CAHmom2	c6-4 ; c9-1
CDH	c7 ; c9-1
CENMOB	c5-3 ; c9-1
Centre de gravité	c5 ; c6-1 ; c9-2 ; c9-3
Centres mobiles (Voir agrégation)	
Chi-deux (Voir Khi-deux)	
Construction ascendante hiérarchique	c4 ; c9 ; c10
Construction descendante hiérarchique	c7
Contributions	c8-2
Corrélation	
De Bravais-Pearson	a1-3
De rangs (Spearman)	a1-3.1 ; b
De rangs (Kendall)	a1-3.1 ; b
Cramèr	c8-2 ; b
CTRHqual	c8-4
CTRHquan	c8-4
CTRPqual	c8-4
CTRPquan	c8-4
De Rham C.	c6-2 ; c9-1.4 ; b
DessArb	c4-3
Diamètre	c4-1
Dichotomies successives	c1 ; c7
Diday E.	c5-1.3 ; b
DisEuc	c3-3
DisKi2	c3-3 ; c9-1.2
DisJac	c3-3 ; c9-1.2
Disjonctif (tableau, voir forme disjonctive)	
Dispersion	c5-1; c6-1
Distance(s)	
De Jaccard	c3-2.2
Du Khi-deux	c3-2.2 ; a1-3

- Euclidienne — c3-1 ; c3-3 ; a1-3
- Indices de distances — a1
- Recalcul des distances — c4-1
- Ultramétrique (Voir ultramétrique)

Effet de chaîne — c4-1.3
Fages R. — c7-3.3 ; b
Forme disjonctive complète (données sous) — c3-1
Formes fortes — c5-1.3 ; c5-2.2
Foucart T. — c1 ; c3-1.2 ; b
Guinochet M. — c2-2 ; b
Guittonneau G-G. — c3-1.2
Guttman L.(effet) — c2-1
Heuristique — c1-2
Hiérarchie

- Construction ascendante — c4
- Construction descendante — c7
- Dessin — c4-3
- Indicée — a2-1
- Interprétation — c8-2
- Troncature — c9-2.1 ; c9-4 ; c10.1

Hubert L. — c7-3.2 ; b
Huyghens C. — c5-1.2 ; b
Indices de distances — a1
Indices de similitude — a1
Informatique — c1-2 ; c3-1.2
Interprétation (aides) — c8 ; c10-1
Inversion (dans une hiérarchie) — c4-1 ; c7-4
Jaccard P. — c3-1.3 ; a1-2 ; b
Jambu M. — c4-1 ; c6-1 ; b
Jardine N. — a2-2 ; b
Khi-deux (ou Khi-carré) — c3-1 ; c3-2 ; c7-2.1 ; c8-3 ; c9-2.3 ; a1-3.2
Lambert J.M. — c7 ; c8-2
Lance G.N. — a1-3.2 ; b
Lerman I.C. — a1-4 ; b
Linné — c1-2
Métrique (= distance, voir ce mot)
Moment d'ordre deux — c5-1 ; c6 ; c8-1
Moment inter-classe — c5-1.2 ; c6 ; c8-1
Moment intra-classe — c5-1.2 ; c6 ; c8-1
Niveau d'agrégation — c1 ; c4 ; c6-2 ; c9-2.1
Nœud — c1 ; c4 ; c6-3 ; c7-2.2
Nuées dynamiques (Voir Agrégations autour de Centres mobiles)
Ordonnance — a1-1 ; a2-1.1
Ordre (sur les distances) — a2-2.1
Partition

- Choix d'une partition initiale — c5-1 ; c5-2.1
- Interprétation — c8-1.1 ; c8-2.1
- Obtenue par troncature — c9-2.1 ; c10-1
- Recherche d'une partition — c5-1

Phi — a1-2
PHYTOS (exemple de données) — c2-2 ; c3-2.2 ; c4-2.2 ; c7-5.2 ; c8-3.2
Phytosociologie — c2-2 ; c3-1.3 ; c3-2.2 ; b
Pondération des distances — c4-1.1
Psychologie — c1-4 ; c2-1
PSYSOC (exemple de données) — c2-1 ; c3-1.1 ; c3-2.1 ; c4-2.1 ; c5-2 ; c6-3 ; c7-5.1 ; c8-3.1
Recalcul des distances — c4-1.1 ; c4-1.2 ; c6-1 ; c6-2 ; c7-4 ; a2-2.2
Roux G. — c2-2 ; a1-2;1 ; b

Roux M. c3-1.2 ; c7-3.3 ; a1-2.2 ; a1-2.2 ; a2-1.2 ; b
Segmentation c1-4 ; b
sélection
    d'objets c7-3
    de variables c7-2.1
Sibson R. a2-2 ; b
Sneath P.H.A. c1-2 ; a1-2 ; a1-2.1 ; a1-2.2 ; b
Sokal R.R. c1-2 ; a1-2 ; a1-2.1 ; a1-2.2 ; b
Taxinomie c1-4 ; c9-3 ; c10-1 ; b
Todd E. c2-1 ; c3-2.1 ; b
Transposition (d'un tableau) c3-3
TRONCAT c9-4
Troncature c1-1 ; c9-2.1 ; c9-4 ; c10-1
Typologie c1-4 ; c9-3 ; b
Ultramétrique a2
Variables
    Rôle des variables c8
    Pondérations des variables a1-3.2
Voisins réciproques c6-2 ; c10-2 ; b
Volle M. c2-1 ; c3-1.2 ; b
Ward J.H. c6-1
    Méthode de Ward : voir agrégation par le moment d'ordre 2
Williams W.T. c7-2.1 ; a1-3.2 ; b